VOL 4 NO 2 | SPRING 2024

International Journal of Law, Ethics, and Technology

La Nouvelle Jeunesse

THE INTERNATIONAL JOURNAL OF LAW, ETHICS, AND TECHNOLOGY

SPRING 2024

La Nouvelle Jeunesse

Publication: The International Journal of Law, Ethics, and Technology

cite as Int'l J. L. Ethics Tech.

ISSN: 2769-7150 (Online) | 2769-7142 (Print)

Publisher: La Nouvelle Jeunesse

Address: 655 15th Street NW, Washington, DC 20005

Date: May 28, 2024

THE INTERNATIONAL JOURNAL OF LAW, ETHICS, AND TECHNOLOGY assumes a paramount role as a dynamic and intellectually stimulating platform dedicated to the meticulous exploration of the intricate interplay between technology, ethics, and the law. As a distinguished peer-reviewed publication, we aim to highlight emerging legal issues by prioritizing exceptional original scholarship that traverses diverse academic disciplines. In our unwavering pursuit of academic excellence, we actively foster contributions that exhibit profound depth and insightful analyses within doctrinal and critical frameworks. Moreover, we enthusiastically embrace interdisciplinary research endeavors that aim to unveil the multifaceted dimensions of the law, drawing upon the diverse perspectives offered by the social sciences and humanities. Rather than regarding law, ethics, and technology as distinct and isolated realms, our journal proudly stands as a nurturing ecosystem that fosters a dynamic and inclusive dialogue. Through a holistic amalgamation of these traditionally delineated fields, we strive relentlessly to engender a comprehensive understanding of the ever-evolving contemporary society we find ourselves in. With open arms and genuine enthusiasm, we sincerely invite scholars from every corner of the globe, urging them to contribute their invaluable knowledge and expertise to this vibrant and intellectually stimulating forum of global knowledge exchange.

SUBSCRIPTIONS: The print version of the *International Journal of Law, Ethics and Technology* is available to individuals and institutions as pre the approval by the editors. To request a place on the list, please email us at info@ijelt.org.

SUBMISSIONS: Please send articles, responses, letters to the editors, and anything else we ought to consider for publication to the *International Journal of Law, Ethics, and Technology* at submissions@ijlet.org.

CORRESPONDENCE: Please write to the *International Journal of Law, Ethics, and Technology* at info@ijlet.org.

Table of Contents

This page intentionally left blank

LAW AND ECONOMICS IN SURROGACY MARKETS

Elizabeth Tharakan*

Abstract: The concept of reproductive justice can offer a framework for complex and nuanced analyses of surrogacy, taking into account the agency of surrogates and potential vulnerability of intended parents. When we define surrogacy contracts, we mean the situation in which the fertilized egg of a married couple is inserted into the uterus of a surrogate mother and carried to term by a woman who has no genetic connection with the fetus. Surrogacy contracts offer a way to supply a genetic connection that adoption contracts cannot provide. This paper argues that surrogacy with heavy regulation is a step forward into the future because it incorporates the best legal solution to custody disputes while also attaining the most economically efficient solution. This paper argues the same about surrogacy markets becoming a standard and prudent practice because calling babies a Pareto-efficient exchange that benefits everyone at the expense of no one. Surrogacy can be an excellent option for a set of biological parents who desire a child but cannot carry a pregnancy for health reasons, with potential framework for the psychological and sociological effects of mental health and abortion. The status of surrogacy agencies within the United States allows for heavy legal and economic regulation of surrogacy markets. When the law of contracts and property are incorporated into surrogacy markets, it is easier and more "efficient" for couples to make bargains with surrogate gestational carriers.

Keywords: Economics; Surrogate; Surrogacy; Gestation; Pregnancy

* Southern Illinois University Carbondale, United States.

Table of Contents

INTRODUCTION TO SURROGACY

Baby Mama is a 2008 movie in which Tina Fey plays a successful businesswoman who discovers that she is infertile and hires a working-class woman to be her surrogate. Surrogacy may seem like a modern phenomenon, but it actually has ancient biblical origins. In the book of Genesis, when Abraham's wife Sarah discovered that she was barren, she had her slave girl, Hagar, carry Abraham's child.[1] However, surrogacy markets are heavily regulated because surrogacy is controversial. When we define surrogacy contracts, we mean the situation in which the fertilized egg of a married couple is inserted into the uterus of a surrogate mother and carried to term by a woman who has no genetic connection with the fetus. Adding new members to a family without procreation comes with all sorts of legal and ethical conundrums. Surrogacy contracts offer a way to supply a genetic connection that adoption contracts cannot provide. This paper argues that surrogacy with heavy regulation is a step forward into the future because it incorporates the best legal solution to custody disputes while also attaining the most economically efficient solution.

It is easier for a woman to avoid pregnancy by using a surrogate if she has a medical reason for wanting one. For example, she could be diabetic, she could be aged, or there could be something wrong with her uterine tract. These medical factors would prevent the genetic mother from carrying a pregnancy on her own. For example, a diabetic woman with uncontrolled blood sugar or a hemoglobin A1C count of greater than 6.0 might cause birth defects to her baby.

I. GLOBAL SURROGACY MARKETS

Surrogacy is an evolving field because the process is becoming legal in a variety of states and legal protections for it are growing. Gestational surrogacy (via the in vitro fertilization method) is an expensive operation. The intended father's sperm fertilizes the intended mother's eggs, and the fertilized egg is subsequently transferred into the surrogate's uterus. The appropriate preparation of the surrogate mother involves hormones, pills, and significant lifestyle changes as prescribed by doctors. Surrogacy is more widespread, costing at least $15,000 for a novice surrogate mother and the price can go up to $25,000 for an experienced surrogate.[2]

The legal status of surrogacy changes across countries and regions. Some countries forbid it while others allow for only altruistic and not commercial surrogacy. Finally, in many countries, legal surrogacy is neither expressly prohibited nor permitted.[3] Because sometimes it is cheaper or easier to find an international surrogate than a domestic one, agencies often sponsor international surrogacy because they have access to untapped markets. International surrogates, specifically, are willing to carry pregnancies for much lower costs.

The surrogacy market in India is flourishing currently. The Indian surrogacy market is valued at over two billion dollars. (Shetty at 1633) Gay male couples and infertile heterosexual couples flocked to India because their own countries prohibited

---

[1] Genesis 16:1-2, King James Version.

[2] Hatzis, A. N. (2003). Just the oven: a law and economics approach to gestational surrogacy contracts. *Perspective for the Unification or Harmonisation of Family Law in Europe. Antwerp: Intersentia*.

[3] Hevia, M. (2018). Surrogacy, privacy, and the American Convention on Human Rights. *Journal of Law and the Biosciences*, *5*(2), 375-397.

surrogacy or because it was cheaper in India. India is an ideal host country for surrogacy-it has fancy medical tourism facilities, English-speaking medical professionals, and a relatively large supply of poor women willing to provide gestational care. Surrogates are often women who do not speak English, live in slums, or are squatters. The Indian government does not monitor surrogacy, but medical professionals do.

Many fertility specialists are involved in recruiting potential surrogates through agents and brokers. Once the surrogate is identified, some fertility specialists even hire intermediaries to manage houses in which surrogates are required to stay for a few days after the embryo transfer and during the later term of the pregnancy. Surrogates are sometimes required to remain in these homes for the duration of their pregnancies. In many of those residences, surrogates are not allowed to leave. Doctors claim that requiring surrogates to live together in one place is the only way to ensure that women receive adequate nutrition, take vitamins, and avoid strenuous exercise.

Indeed, anti-commodification adherents in India argue surrogates only enter that status because of their poverty. Some people in India focus on surrogacy as an example of the problems of the capitalist system more broadly. For example, the feminist group All India Democratic Women's Association (AIDWA), which is the women's wing of the Communist Party of India (Marxist), believes that compensated surrogacy should be banned because it exploits surrogates. Specifically, poor women turn to being surrogates and carrying pregnancies for women of means because the poor women are desperate for money and compensation.

In the United States, malpractice lawsuits are prolific. By contrast, Indian professionals do not operate in a similar environment because litigation is not common. First, access to justice in India is much more difficult for poor surrogates. This is, in part, because lawyers in India cannot base their fees on the outcome of the litigation because the Bar Council of India prohibits it. , Whereas in America poor litigants can get contingency-fee lawyers. This essentially means that poor people who cannot afford lawyers' fees will never be able to bring a lawsuit. Indian courts also fail to award high punitive damage awards. Any medical negligence claim must instead be brought to consumer protection courts pursuant to the Indian Consumer Protection Act.

II. SURROGACY CASE STUDY

The case study of celebrity couple Chrissy Teigen and John Legend demonstrates how Teigen was pregnant with her third child a few months before a surrogate named Alexandra became pregnant with Teigen’s fourth child. Teigen explained in an interview that the couple learned that Alexandra was pregnant with Wren "as we crept toward the safe zone of my own pregnancy" with daughter Esti, now 5 months. "We ate hot pot to celebrate, watched *Vanderpump Rule* with our growing bellies, our families blending into one for the past year," she recalled. "Just minutes before midnight on June 19th, I got to witness the most beautiful woman, my friend, our surrogate, give birth amidst a bit of chaos, but with strength and pure joy and love." (Portee).

“But for me, I think the way I operate with anybody in our house, whether it’s nannies or my mother living with us, or my friends that are in the house or security or whatever, I want an atmosphere where everybody feels really comfortable in our home and that extended to her,” she explained. “I wanted her to feel like she could take off

her shoes and kick up her feet on our couch. We could watch TV together and my daughter could play with her daughter up in her room. It felt like she could come over anytime, and I feel like we do still have that relationship. It's been really wonderful." (Id.)

The couple's birth announcement chose to name their son after their surrogate. "We want to say thank you for the incredible gift you have given us, Alexandra," she wrote in the post. "And we are so happy to tell the world he is here, with a name forever connected to you, Wren Alexander Stephens." (Id.)

III. THE SURROGACY CONTRACT

Should surrogacy contracts be regulated or prohibited by the state? How are ownership rights established in surrogacy when there is a dispute over who gets the baby? A preliminary review of the limited literature available on this novel topic demonstrates that scholars are mostly in favor of surrogacy contracts.

The surrogate mother must "obey all doctor's orders made in the interests of the child's health. These orders could include forcing her to give up her job, travel plans, and recreational activities. The doctor could confine her to bed, regulate her diet rigidly, and order her to submit to surgery and to take drugs."[4] (Epstein). To be sure, the surrogate mother has to surrender some autonomy to the biological parents. Such arise the terms of the contract, with the hope for greater gains for both the surrogate mother and the biological parents. These restrictions are so burdensome that some people would not enter into a surrogacy transaction at all. However, for some, the cost of these additional restrictions is perceived as being low enough to be worthwhile. "To argue that these contractual terms are inconsistent with the autonomy of the surrogate mother is to miss the function of all contractual arrangements over labor," argued Epstein.[5]

When contract and property law is invoked, theories of private property come into play. But with the market for babies, we must incorporate family law as well because babies are their own agents and we value human life and do not treat it as property. Lawrence Gostin argues for a "best interests of the child" standard because babies cannot be bought: "Judge Posner thinks about surrogacy arrangements in terms of economic liberty: The parties are in relatively free and equal bargaining positions, the arrangements are mutually beneficial, and third parties (notably the children) are not harmed."[6] (Gostin). According to Gostin's interpretation of Judge Posner's reasoning, surrogacy contracts yield a desirable win-win result, so they should be allowed.

In recent history, surrogacy contracts have been upheld. In 1987, an American court heard the case of In Re Baby M. Through the Infertility Center of New York, a surrogate and her husband gave birth to Sarah Elizabeth Whitehead and listed themselves as the birth parents on the birth certificate. When the biological parents, William Stern and his wife, named the child Melissa Stern or Baby M, they paid the

---

[4] Epstein, R. A. (1995). Surrogacy: the case for full contractual enforcement. *Virginia Law Review*, 2334.

[5] Ibid., 2335.

[6] Gostin, L. O. (2001). Surrogacy from the Perspectives of Economic and Civil Liberties. , available at chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://scholarship.law.georgetown.edu/cgi/viewcontent.cgi?article=2833&context=facpub.

surrogate $10,000. The New Jersey Supreme Court found the surrogacy contract over Baby M illegal and invalid.[7] (McEwen). Then in 1989, two bills prohibiting surrogacy failed to pass in Congress. (McEwen). Then in the 1990 California case, Johnson v. Calvert, the surrogate wanted to keep the child but a court found the genetic mother to be the true mother.[8] (McEwen). In essence, the court honored the surrogacy contract with specific performance of the contract rather than money damages as the form of relief. Without legally enforceable contracts, courts may have come to a different conclusion.

IV. SURROGACY AND ABORTION

If the infant is found to be seriously defective during the surrogate's pregnancy, can the parents order an abortion? One possible response is that the surrogate gestational carrier would refuse to accept this particular contract term, at which point someone has to decide who has custodial obligations to the child when it is born. "It is the father who has, under contract, the long-term obligations for the child, and it cannot be regarded as unjust or unwise that his https://www.usccb.org/news/2024/surrogacy-injustice-all-involved-bishop-barron-says-support-pope-francisdecision should determine whether the abortion should take place for precisely those reasons that are so important to ordinary married couples"[9] (Epstein.)

Regulating abortions comes with the thorny territory of American constitutional jurisprudence, such as Roe v. Wade and Casey v. Planned Parenthood, but regulating abortion within the context of surrogacy is something that must be done if we are to make surrogacy a valid option for people within the United States.

Another abortion problem that might occur is that the surrogate mother develops some sort of health complication in response to the pregnancy and wants to abort the fetus, but the biological parents want their child. There is really no ethical solution to this dilemma, but the law over the past twenty-five years allowed the surrogate mother to have an abortion because it's her body.

The law becomes even thornier in the wake of the Dobbs v. Jackson Women's Health Organization decision, because it left the abortion question up to the states. For example, abortion is legal in New York but illegal in Texas. Thus, whether a surrogate could or should abort the fetus if she wanted or needed to would depend on which state she lived in and how the state regulated those laws. The surrogate would have different rights and obligations vis-à-vis the fetus in those two states.

V. SURROGACY AND PSYCHOLOGY

Another reason that society needs legally enforceable contracts is that the surrogate may form emotional attachments to either the child or the intended parents. There is a close relationship between the child's biological parents and the surrogate. When the surrogate gives birth, this relationship ends.

---

[7] Ibid.
[8] McEwen, A. G. (1999). So You're Having Another Woman's Baby: Economics and Exploitation in Gestational Surrogacy. *Vand. J. Transnat'l L.*, *32*, 271.
[9] Epstein, 2336.

The surrogate may also face other psychological issues, such as postpartum depression or stress and anxiety. She may worry about the financial burdens that she faces in her everyday life, which motivated her into being a surrogate for hire. Pregnancy alone can cause things like depression to get worse. Hormones and the surrogacy environment might affect any given mental illness.

Psychologists administer the Personality Assessment Inventory (PAI) to would-be surrogates. The test assesses personality and psychopathology and is divided into 22 scales, including anxiety-related disorders, aggression, alcohol and drug problems, and more. The evaluation also requires you to speak with a psychologist, usually from the surrogacy agency. The psychologist will ask things such as why the prospective childbearer wanted to become a surrogate, aspects of the surrogacy journey, the support system, etc. The psychologist may also want to talk to the surrogate's partner or spouse.

How a couple can do right by a surrogate is an ethical question requiring psychological understanding. Surrogacy counselors or psychologists can play a crucial role in helping intended parents and surrogates understand and cope with the psychological impact of surrogacy. They provide support, education, and other guidance because they have skills and knowledge in helping people deal with their feelings.

Susan Golombok, a professor of family research and director of the Centre for Family Research at the University of Cambridge, led a team of British researchers to find that children born with the help of a surrogate may have more adjustment problems than those born to their mother via donated eggs and sperm. Their results, published in the June issue of the Journal of Child Psychology and Psychiatry, suggest that it's difficult for kids to deal with the idea that they grew in an unrelated woman's womb.

VI. SURROGACY AND THE CATHOLIC CHURCH

Another objection to surrogacy comes from Pope Francis, the head of the Roman Catholic Church, who recently denounced surrogacy in the news. Bishop Robert Barron, speaking in defense of the pontiff's views, stated the following: "The commercialization of women and children in surrogacy is underlined by the belief that there is a right to have a child. The child becomes an object for the fulfillment of one's desires instead of a person to be cherished. In this way, the genuine right of the child to be conceived through the love of his or her parents is overlooked in favor of 'the right to have a child by any means necessary.' We must avoid this way of thinking and answer the call to respect human life, beginning with the unborn child."[10] (Barron)

Barron added, "It might be the case that couples earnestly want to have children without resorting to surrogacy, but painful and even life-threatening medical obstacles make childbirth hazardous or impossible. The serious prospect of a life without biological children has been dismissed by some, but we have a responsibility to accompany these couples in their suffering. The Church teaches that married couples are not obliged to actually have children, but to be open to any life that might be the

[10] *Surrogacy is an Injustice to All Involved, Bishop Barron Says in Support of Pope Francis*, United States of Conference of Catholic Bishops (January 10, 2024), https://www.usccb.org/news/2024/surrogacy-injustice-all-involved-bishop-barron-says-support-pope-francis.

fruit of their union. The desire to utilize surrogacy might feel like the desire to form a family naturally, but no matter how well-intentioned, surrogacy always does grave injustice to the child, any discarded embryos (who are our fellow human beings), the commodified birth mother, and the loving union of the spouses." This consideration of the discarded embryos is based on the fact that, according to Catholic teaching, an embryo is alive because life begins when the egg and the sperm meet at conception.

VII. LAW AND ECONOMICS

Law and economics is an emerging field that transitions us from the existing literature of "what is" to the idea of "what should be" or "what could be" in an ideal world. The idea is that we can use efficiency to find the best outcome. Law and economics scholarship employs two different kinds of analysis, positive (descriptive measures of what is) and normative (an "ought" analysis of what should be). Normative analysis explains what should happen, which is that surrogacy markets should be heavily regulated and surrogacy contracts should come with the force of law. With respect to positive economic analysis of legal issues, the question is that if this policy is adopted, what predictions can we make about the likely economic impacts?

In predicting these behavioral responses, a positive analysis will assume that most individuals are motivated by rational self-interest to maximize utility. But with surrogacy, legislators, executives, and judges measure the interests of not just the bargaining parties but also that of the third party, the baby. Pareto efficiency would ask of any transaction or policy: will this transaction make somebody better off while making no one worse off? Kaldor-Hicks efficiency, by contrast, would ask whether this collective decision would in terms of cost benefit analysis generate sufficient gains so that they could compensate the losers sufficiently to render them indifferent to it but also have gains left over for themselves? (Trebilcock and Keshvani).[11] Positive analysis explains what happens with contract and property law in surrogacy disputes.

Economic analysis of contract law in particular has offered a theory on which promises should be enforced. Under this approach, a contract should be enforced when it makes two people better off, without making anyone worse off.[12] This is known as a Pareto-efficient exchange. Surrogacy contracts meet the criteria of improving everyone's welfare without harming anyone, but there is a commodification argument that says that women shouldn't be able to trade their inalienable parental rights for money. The argument says that the value of giving up an emotional attachment to the child and the costs of laboring are less than the value of being compensated for carrying the child. However, the Pareto-efficiency argument would respond that carrying the child and compensation for it does not make anyone worse off. The only people who would disagree are highly moralistic thinkers.

Making promises requires deferred exchange rather than instantaneous exchange. If exchange were instantaneous, there would be no reason for promises. The agency game demonstrates why we need contracts to buttress these promises: In this game, the first player might be an investor in a business, a consumer buying goods, a bank account holder, or the purchaser of an insurance policy. If the first player hands over

---

[11] Trebilcock, M. J., & Keshvani, R. (1991). The role of private ordering in family law: a law and economics perspective. *U. Toronto LJ*, *41*, 533.
[12] Ibid., 283.

the asset to within the second player's control, "the second player decides whether to cooperate or appropriate. Cooperation is productive, whereas appropriation is redistributive." Productivity could take the form of realized gains such as the profit from investment, the surplus from trade, or the interest from a loan. The parties divide the product of cooperation between them, so both of them benefit from having played the game or made the contract. By contrast, appropriation redistributes the asset from the first player to the second player – this is the result of "finking."

Without contracts to enforce promises, people would be inclined to break their promises and not cooperate in order to redistribute funds. An innovator in Silicon Valley might ask a business guru to invest $1 million in a start-up fund to develop a new computer chip. By developing the chip, the innovator can turn $1 million into $2 million. The innovator promises to develop the chip and share the profits of $1 million equally with the investor.[13] Instead of developing the chip, however, the innovator might try to take the investor's $1 million and self-deal. An enforceable promise to develop the chip will prevent the innovator from appropriating the money; so, the investor will trust the innovator and invest the money. Contract law makes promises more trustworthy.

Commitment is achieved by foreclosing the opportunity to appropriate (to take the money and run). The opportunity to appropriate is foreclosed by the high price of liability for breach. A commitment is credible when the other party observes the foreclosing of the appropriation opportunity. In terms of surrogacy contracts, surrogate mothers might be inclined not to give over the babies to which they gave birth. Therefore, surrogacy invokes the law of contracts to be a sustainable process. When prospective parents find a surrogate through an accredited agency, they sign a contract that is backed up by a credible commitment.

We can think about reproductive justice in terms of rights (where a right is a "just claim," as the philosophical definition goes) and their protections. To preserve rights, we can start by using property law as a tool. Aristotle espouses a system of private property: "Property should be in a certain sense common, but, as a general rule, private; for, when everyone has a distinct interest, men will not complain of one another, and they will make more progress, because everyone will be attending to his own business" (Aristotle in Epstein.)[14] Private property comes with the labor theory of value, where the man who goes to pick the apple will be considered as owning it and having the right to eat it because he did the work for it and his property interest should properly be called a right.

Law is unnecessary and undesirable where bargaining succeeds, but necessary and desirable where bargaining fails. (Cooter). [15] Bargaining occurs through communication, but communication comes with certain costs. These include renting a conference room, hiring a secretary or notetaker and clearing schedules to set up a meeting time. These costs are called transaction costs.

---

[13] Ibid., 285.
[14] Epstein, R. A. (1995). Surrogacy: the case for full contractual enforcement. *Virginia Law Review*, 2305-2341.
[15] Cooter, 81.

Should babies count as private property of their genetic parents or of the surrogate mother? The answer is that babies, because they are human beings and because we no longer allow ownership of human beings in this country, are their own agents and so courts should use the "best interests of the child" standard common in family law cases. When courts face a custody dispute between the biological parents and the surrogate mother, they should examine factors like financial stability, emotional health and household composition to see who can provide a better home for the child.

It's not that fairness requires the party who causes harm to pay for it, which is often the legal standard, but rather a question of efficiency that the damages be least, which is the economics standard. The Coase theorem sums it up: "When transaction costs are zero, an efficient use of resources results from private bargaining, regardless of the legal assignment of property rights."[16] A corollary to the Coase theorem says when transaction costs are zero, an efficient use of resources results from private bargaining, regardless of the legal assignment of property rights.[17] In the case of surrogacy, transaction costs are not zero but actually high because it costs around $20,000 to hire a surrogate in America and agencies have to act as an intermediary by performing background checks, finalizing the in vitro fertilization treatments with hospitals and doctors, and putting together a surrogacy contract. So according to Coase theorem and its corollary, a private bargain for a surrogate to carry a fetus does not necessarily result in the most economically efficient outcome because transaction costs are high. Rather, there should be some governmental regulation of surrogacy markets.

Damages are more efficient than injunctions as the remedy to a property breach when transaction costs are high. Cooter gives the example of laundry and the electric company. If damages perfectly compensate the laundry when the electric company pollutes, assuming no transaction costs, its profits are the same. An injunction by contrast forces the electric company to abate and not pay damages, so the electric company has no choice in the matter. When transaction costs are high, damages are a better remedy. When transaction costs are low, injunctions are a better remedy. Injunctions and other kinds of specific performance would not apply to surrogacy contracts because transactional costs are high in surrogacy markets, where damages would be the better remedy. So even though courts have historically provided specific performance, a form of injunctive relief, in surrogacy contracts by requiring that the surrogate gestational carrier hand over the baby to the genetic parents, Coase theorem would suggest that damages yield a more efficient outcome. In this way, courts could require that if a surrogate mother got attached to the child and wanted to keep it, she could as long as she repaid the genetic parents with the $15,000 they were going to pay her for the child. However, Coase theorem is limited in its applicability because it does not recognize moral externalities involved in letting a surrogate mother keep the child. For example, a surrogate mother couldn't provide a blood transfusion to her child because she would not necessarily have the same blood type. That's one biological problem with the "efficiency" solution of damages and courts not awarding specific performance. This paper argues that governments should therefore regulate surrogacy contracts with specific performance rather than with damages, and thus these contracts would have legal weight.

---

[16] Ibid.
[17] Ibid., 85.

When courts can remedy the victim of a broken promise, contract law is enforceable. “People continually make promises: sales people promise happiness; lovers promise marriage; generals promise victory; and children promise to behave better. The law becomes involved when someone seeks to have a promise enforced by the state.”[18] The bargain theory of contracts says there needs to be a bargained-for exchange, also known in legal terminology as consideration. There must be a quid pro quo (consideration), as well as an offer and an acceptance. In surrogacy contracts, a healthy child is exchanged for money to carry a pregnancy.

CONCLUSION

The emerging technology involved in surrogacy is expanding rapidly: it will eventually be more cost-effective such that surrogacy becomes commonplace for infertile and LGBTQ couples. Life insurance was once viewed as a form of trafficking in human life. But as traditional understandings of law evolved into more modern ideas, life insurance became a standard and prudent practice. The same reasoning applies to surrogacy, that it was once seen as wrong but is now seen as a good practice.

A framework of legal regulation could thwart some of the harms of exploitation that could happen in American surrogacy markets. We have used the example of the Indian surrogacy market to show that surrogacy could be easier than it is when laws step in to regulate humans. The Founding Fathers of America said, “If men were angels, no government would be necessary.” In this way, government and the future legal landscape could step in to protect all of the stakeholders in surrogacy markets, including the surrogate, the biological parents, the lawyers, the doctors, the agencies, and the fertility specialists.

The concept of reproductive justice can offer a framework for complex and nuanced analyses of surrogacy, taking into account the agency of surrogates and potential vulnerability of intended parents. This paper argues the same about surrogacy markets becoming a standard and prudent practice because calling babies a Pareto-efficient exchange that benefits everyone at the expense of no one. Surrogacy can be an excellent option for a set of biological parents who desire a child but cannot carry a pregnancy for health reasons, with potential framework for the psychological and sociological effects of mental health and abortion. The status of surrogacy agencies within the United States allow for heavy legal and economic regulation of surrogacy markets. When the law of contracts and property are incorporated into surrogacy markets, it is easier and more “efficient” for couples to make bargains with surrogate gestational carriers but the Coase theorem’s efficient solution of awarding damages rather than specific performance when surrogacy contracts are breached does not work with existing legal structures at enforcing contracts. This paper would invite skeptics of surrogacy markets to keep up with modern times.

[18] Ibid., 276.

BIBLIOGRAPHY

1. Barron, Robert. (Jan. 10, 2024). United States Conference of Catholic Bishops Office of Public Affairs Statement, available at https://www.usccb.org/news/2024/surrogacy-injustice-all-involved-bishop-barron-says-support-pope-francis.

2. Carroll, Linda. (2013). New study tracks emotional health of surrogate kids, available at: https://www.today.com/health/new-study-tracks-emotional-health-surrogate-kids-6c10366818.

3. Cooter, R., & Ulen, T. (1988). Law and economics.

4. Epstein, R. A. (1995). Surrogacy: the case for full contractual enforcement. *Virginia Law Review*, 2305-2341.

5. Gostin, L. O. (2001). Surrogacy from the Perspectives of Economic and Civil Liberties, available at chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://scholarship.law.georgetown.edu/cgi/viewcontent.cgi?article=2833&context=facpub.

6. Hatzis, A. N. (2003). Just the oven: a law and economics approach to gestational surrogacy contracts. *Perspective for the Unification or Harmonisation of Family Law in Europe. Antwerp: Intersentia*.

7. Hevia, M. (2018). Surrogacy, privacy, and the American Convention on Human Rights. *Journal of Law and the Biosciences*, *5*(2), 375-397.

8. Kalantry, Sital (2018) "Regulating Markets for Gestational Care: Comparative Perspectives on Surrogacy in the United States and India,"Cornell Journal of Law and Public Policy: Vol. 27 : Iss. 3 , Article 8. Available at:https://scholarship.law.cornell.edu/cjlpp/vol27/iss3/8.

9. King James Version Bible, 2022 online version at Bible Gateway.

10. McEwen, A. G. (1999). So You're Having Another Woman's Baby: Economics and Exploitation in Gestational Surrogacy. *Vand. J. Transnat'l L.*, *32*, 271.

11. Trebilcock, M. J., & Keshvani, R. (1991). The role of private ordering in family law: a law and economics perspective. *U. Toronto LJ*, *41*, 533.

THE EVOLUTION OF THE INCENTIVES FOR ANTI-CORRUPTION CORPORATE COMPLIANCE PROGRAMS IN THE INTERNATIONAL LEGAL ORDER

Dalila Martins Viol*

Abstract: The corporate compliance programs have proliferated worldwide in the last two decades. This innovative study explores the role of the International Anti-Corruption Regime (IACR) in this phenomenon. Analyzing documents from 18 international actors, it identifies 52 directly promoting compliance programs, beginning in 2002, after the regime's 1996 inception. These promotions are primarily in non-binding instruments, which present compliance programs as an anti-corruption strategy for business, government, and collective actions. This mapping reveals a transition in the IACR's inception of these programs, starting from a bribery-focused to a broader anti-corruption approach, and more recently endorsing a connection with the Environmental, Social, and Governance (ESG) agenda. This article also offers insights about the link between the IACR and countries' legal reforms promoting compliance. Shedding light on compliance program evolution within the IACR, the study contributes to understanding the strategy's rise and to an emerging legal scholarship in international compliance studies.

Keywords: International Law; Anti-Corruption; Corporate Compliance Programs; Legal Incentives

* School of Law of the Getulio Vargas Foundation, São Paulo, Brazil.

Table of Contents

INTRODUCTION

While corporate compliance was a topic of interest in a limited number of countries in the 2010s, nowadays, anti-corruption systems are adopted in companies worldwide.[1] The increase in the number and importance of corporate compliance programs is related to a phenomenon that the literature recognizes as the "era of compliance."[2] In general lines, compliance programs are organization's internal systems and procedures for helping to ensure that the organization – and those working there – comply with legal requirements and internal policies and procedures.[3] While scholars and policymakers have highlighted the spread of corporate compliance programs, they have not devoted sufficient attention to the influence of legal drivers of this diffusion. Commentators have assumed that the development of international anti-corruption conventions, along with subsequent domestic regulations that establish liability or impose more severe penalties on legal persons involved in corruption, has led companies to implement anti-corruption measures.[4]

It is indisputable that the International Anti-Corruption Regime (hereinafter "IACR") has played a crucial role in coordinating states' domestic responses to corruption, influencing whether and how states regulate it.[5] Since the start of the twenty-first century, numerous states have enacted anti-corruption statutes and regulations, driving several legal reforms around the world.[6] This framework serves as a significant incentive for companies to establish compliance programs, reducing the risk of sanctions for corruption.[7] However, there is more to the story.

Seeking insights into the treatment of anti-corruption corporate compliance programs (hereinafter "compliance programs") within the IACR, I mapped over one hundred documents from 18 international actors, aiming to contribute to an understanding of the factors that have positioned this strategy at the core of anti-corruption policies for both businesses and governments around the world.[8] This article reveals that direct incentives for compliance programs are notably absent in international anti-corruption conventions. Instead, references to compliance programs

---

[1] OECD. (2020). *Corporate Anti-Corruption Compliance Drivers, Mechanisms, and Ideas for Change*. www.oecd.org/daf/anti-bribery/Corporate-anti-corruption-compliance-drivers-mechanisms-and-ideas-for-change.pdf.

[2] *See, e.g.,* Richard S. Gruner, *General Counsel in an Era of Compliance Programs and Corporate Self-Policing*, 46 EMORY LAW JOURNAL 1113 (1997); Sean. J. Griffith, *Corporate Governance in an Era of Compliance*, 57 WILLIAM AND MARY LAW REVIEW 2075 (2016); Robert C. Bird & Stephen K. Park, *The Domains of Corporate Counsel in an Era of Compliance*, 53 AMERICAN BUSINESS LAW JOURNAL 203 (2016); Rory Van Loo, *Regulatory Monitors: Policing Firms in the Compliance Era*, 119 COLUMBIA LAW REVIEW 369 (2019); Asaf Eckstein, *The Virtue of Common Ownership in an Era of Corporate Compliance*, 105 IOWA LAW REVIEW 507 (2020).

[3] SFO. *Evaluating a Compliance Programme*. https://www.sfo.gov.uk/publications/guidance-policy-and-protocols/guidance-for-corporates/evaluating-a-compliance-programme/#_ftn3, on 20 Jan. 2024.

[4] *See, e.g.,* Kevin E. Davis & Veronica R. Martinez, *Transnational Anti-bribery Law*, in THE CAMBRIDGE HANDBOOK OF COMPLIANCE 924 (Benjamin van Rooij & D. Daniel Sokol ed., 2021).

[5] *See, e.g.,* BARNALI CHOUDHURY & MARTIN PETRIN, CORPORATE DUTIES TO THE PUBLIC (2019).

[6] OECD. (2020). *Corporate Anti-Corruption Compliance Drivers, Mechanisms, and Ideas for Change.* https://www.oecd.org/corruption/corporate-anti-corruption-compliance.htm; MATTESON ELLIS, THE FCPA IN LATIN AMERICA: COMMON CORRUPTION RISKS AND EFFECTIVE COMPLIANCE STRATEGIES FOR THE REGION (2016).

[7] *See*, *e.g.*, Kevin E. Davis & Veronica R. Martinez, *Transnational Anti-bribery Law*, in THE CAMBRIDGE HANDBOOK OF COMPLIANCE 924 (Benjamin van Rooij & D. Daniel Sokol ed., 2021).

[8] *See* section 2.

are observed in non-binding instruments of the IACR, demonstrating a diffusion through soft law mechanisms. The first mention was traced back to a document published by Transparency International (TI) in 2002 that encourages companies to adopt compliance programs.[9] The first founded document that stimulates governments to create incentives for companies to adopt compliance was published by the International Chamber of Commerce (ICC) in 2005.[10] This demonstrates a precursor role of private international initiatives in the promotion of compliance programs, even though the majority of documents found in this study that mention compliance programs are from international organizations.

This study found 52 documents – from international organizations (39%), private international initiatives (35%), intergovernmental initiatives (12%), international financial institutions (2%), or from multiple actors (12%) – that specifically promote compliance programs as a strategy against corruption in the last two decades. These documents offer different justifications for the implementation of such a strategy. Most of those stimulate companies to adopt compliance programs as mechanisms to support them in their fight against corruption (63%), with an emphasis on the idea that companies should oppose corruption not just because it is illegal but also because controlling it is beneficial for businesses. Other documents urged states to establish legal incentives for companies to adopt compliance programs as part of their comprehensive public policy against corruption (19%), often assuming that the state's role in fighting corruption should extend beyond merely penalizing companies for misconduct. Some instruments target both companies and governments (10%). Moreover, there are instruments that described compliance programs as a relevant tool in collective actions (8%),[11] which are initiatives involving both private and public

[9] TI and Social Accountability International. (2002). *Business Principles for Countering Bribery: An Initiative of Transparency International and Social Accountability International.* https://www.newsd.admin.ch/newsd/message/attachments/5465.pdf

[10] ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf.

[11] "[…] there is consensus that Collective Action can take four main forms according to the length and breadth of the involved activities (from longer to shorter-term endeavours, encompassing sectorial or project-specific goals) as well as whether they are of a voluntary nature or involve some form of enforceability or external monitoring: **Anti-Corruption Declarations**: Voluntary, principle-based, ethical public statements and commitments regarding integrity principles that can be fostered by a group of companies or a group of companies jointly with other actors from civil society – e.g., an anti-corruption NGO – and/or the public sector – e.g., an anticorruption agency. **Standard-Setting Initiatives**: Development of specific anti-corruption frameworks and standards tailored to address specific sector problems and weaknesses such as a code of ethics, code of best practices, etc., that are developed with the help of business associations or similar organizations, and that help in standardizing certain integrity policies within a specific sector and align individual members practices. **Capacity-Building Initiatives**: Companies jointly share their know-how, resources and tools from their compliance programmes, and with the help of their compliance practitioners, to offer concrete capacity building and training opportunities for other companies that are part (or not) of their supply and value chains, in particular SMEs, as well as for public officials and organizations, and other practitioners from civil society organizations. The aim of these initiatives is to help create or enhance compliance systems and tools in smaller and/or less resourceful organizations. **Integrity Pacts**: Agreements that involve a higher level of commitment from their members, and that are most commonly used in specific public tenders or bidding for large projects in infrastructure, sports events, for procurement procedures, etc., with the aim of preventing bribery, conflicts of interest, etc. They can incorporate an external monitoring and certification process which can include sanctions in case of non-compliance, from lesser ones to even exclusion from the initiative. These distinct types are not rigid. Certain Collective Action initiatives can mix many elements of the different types at the same time or can

actors aiming to combat corruption in specific sectors, understanding that the effective fight against corruption needs collective engagement. Compliance programs within the IACR can serve as instruments for anti-corruption corporate policies, strategies within anti-corruption regulatory frameworks, or tools in collective actions, illustrating the diverse avenues for their dissemination.

The mapping also demonstrates a change in the approach to the model of compliance programs in these documents. In general, the oldest documents promoted compliance programs focused on bribery, followed by a wave of documents that address corruption in a broader sense. More recently, IACR actors have been publishing documents that stimulate compliance programs aligned with an Environmental, Social, and Governance (ESG) approach. These findings offer valuable insights into the historical development and evolving dynamics of the discourse surrounding compliance programs within the IACR.

This study also uncovers that the IACR began advocating for compliance programs as a strategy against corruption in 2002, a significant period after its establishment marked by the Inter-American Convention Against Corruption (hereinafter referred to as the OAS Convention). The recommendations within the IACR for governments to reform their legal systems to promote corporate compliance programs also came later compared to the paradigmatic legal reform within a domestic framework, which occurred in the United States in 1991. The IACR's push for compliance programs appears to be part of a broader movement towards their proliferation. However, this does not discount the potential of the IACR to influence states in legal reform for compliance. As illustrated in this study, some countries, both in the Global North and the Global South, adopted incentives for promoting corporate compliance programs after the IACR made recommendations in this regard. Future research can delve deeper into the relationship between the IACR and domestic regulations concerning compliance programs.

This study contributes to the understanding of the global rise of compliance programs, taking a step towards bridging a gap in the existing literature. Although the compliance programs has "become a key mechanism to markets, societies, and modes of governance across a variety of public and private domains,"[12] scholars do not have a comprehensive understanding of "what mechanism and intervention play a role in shaping it."[13] Corporate compliance programs have also not been "adequately systematized from a theoretical perspective"[14] and are "largely absent from the mainstream corporate law literature."[15] Furthermore, this study offers an original overview of the IACR compared to previous studies on the international anti-corruption

---

evolve in time from one type to another according to the needs and demands of the involved stakeholders." (WEF. (2020). Agenda for Business Integrity: Collective Action – Community Paper. https://www3.weforum.org/docs/WEF_Agenda_for_Business_Integrity.pdf, at 4).

[12] Régis Bismuth, Jan Dunin-Wasowicz & Philip M. Nichols, *The Transnationalization of Anti-corruption Law: An Introduction and Overview*, in THE TRANSNATIONALIZATION OF ANTI-CORRUPTION LAW 1 (Régis Bismuth, Jan Dunin-Wasowicz & Philip M. Nichols ed., 2021), at ii.

[13] Régis Bismuth, Jan Dunin-Wasowicz & Philip M. Nichols, *The Transnationalization of Anti-corruption Law: An Introduction and Overview*, in THE TRANSNATIONALIZATION OF ANTI-CORRUPTION LAW 1 (Régis Bismuth, Jan Dunin-Wasowicz & Philip M. Nichols ed., 2021), at 2.

[14] Stefano Manacorda & Francesco Centonzo, *Preface*, in CORPORATE COMPLIANCE ON A GLOBAL SCALE (Stefano Manacorda & Francesco Centonzo ed., 2022), at v.

[15] Sean. J. Griffith, *Corporate Governance in an Era of Compliance*, 57 WILLIAM AND MARY LAW REVIEW 2075 (2016), at 2080.

field, as its focuses on a specific and central element of this regime: the compliance programs.

The rest of this article unfolds in four parts. Part 2 addresses the lack of consensus on which actors and instruments are part of the IACR, defining the regime and providing an overview of it, and explains the structure of this article. Part 3 maps how and when the IACR approaches compliance programs. Part 4 analyzes the legal instruments of the IACR that directly promote compliance, revealing that the regime has elected compliance programs as a strategy against corruption and demonstrates the changes in the approach to the model of compliance programs in these documents over time. Part 5 offers concluding remarks and suggests new studies on compliance programs from international and comparative perspectives.

I. THE INTERNATIONAL ANTI-CORRUPTION REGIME (IACR)

The IACR lacks precise boundaries, as it is composed of an uncoordinated network of rules, laws, processes, and norms that operate to control corruption, which constantly changes and grows.[16] Different scholars incorporate different documents into the IACR, and these documents are produced for various types of international actors, each of which often has distinct priorities and strategies when it comes to anti-corruption efforts.[17] The different approaches in the literature on the IACR can be attributed to the significant development in this field over the past two decades. Another contributing factor is the diverse nature of these instruments, encompassing both soft and hard law, for instance, which presents challenges in conducting a comprehensive analysis.

In a paper focused on initiatives by international actors that have been reflected in legal instruments, Jose-Miguel Bello y Villarino describes and analyses the evolution of international legal efforts to combat corruption into four overlapping phases.[18] First, there is the "transborder" period, which is characterized by the expansion of domestic anti-bribery regulations, particularly the Foreign Corrupt Practices Act (FCPA) from United States beyond national borders.[19] This period also covers the OECD Convention, which the author understands as the "global FCPA."[20] The second period is the "international" phase, which encompasses regional and global instruments that address not only bribery but also other types of corruption and their connection to state development.[21] As part of this period, Villarino listed the United Union Convention Against Corruption (hereinafter UN Convention) as a global instrument. As regional instruments, he describes the OAS Convention, enacted in the scope of Organization of

---

[16] Régis Bismuth, Jan Dunin-Wasowicz & Philip M. Nichols, *The Transnationalization of Anti-corruption Law: An Introduction and Overview*, in THE TRANSNATIONALIZATION OF ANTI-CORRUPTION LAW 1 (Régis Bismuth, Jan Dunin-Wasowicz & Philip M. Nichols ed., 2021).

[17] Susan Rose-Ackerman, *The Role of International Actors in Fighting Corruption*, in ANTI-CORRUPTION POLICY: CAN INTERNATIONAL ACTORS PLAY A CONSTRUCTIVE ROLE? 3 (Susan Rose-Ackerman & Paul D. Carrington ed., 2014).

[18] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022).

[19] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 349.

[20] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 351.

[21] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 358.

American States (OAS), the African Union Convention on Preventing and Combating Corruption,[22] and the European Framework.[23] Third, there is the "transnational" period, wherein different actors operating within the boundaries of international law, not necessarily states, develop mechanisms to address corrupt behaviors (or suspicions of corruption) within their domains of business.[24] As part of this phase, the author lists the Financial Action Task Force's (FATF) actions to fight corruption, the World Bank sanctions for corrupt practices, the work of the ICC, and the Revised World Trade Organization Global Procurement Agreement (Revised WTO GPA). The fourth and last phase that Villarino identifies is the "disruptive approach," which sets these new initiatives apart by prioritizing anti-corruption as their primary objective rather than treating it as a secondary or subsidiary goal.[25] He further notes that many of the initiatives of the fourth phase remain unimplemented proposals (such as the human right to live free of corruption or establishing an international anti-corruption court), with the Extractive Industries Transparency Initiative (EITI) being a notable exception.

Prior to Villarino's study, Jan Wouters, Cedric Ryngaert, and Ann Sofie Cloots (WRC) emphasized that the IACR has seen substantial strengthening since the 1990s, with significant but not sufficient progress made at both global and regional levels.[26] In the authors' effort to provide an overview of this regime, they categorize various international anti-corruption instruments into three distinct groups based on the type of instrument. The first category contains the "international anti-corruption instruments," in which the author includes the UN Convention, the European Instruments,[27] and "other regional anti-corruption instruments," encompassing instruments from OECD[28]

---

[22] "Some other minor regional treaties that predate these broader conventions can also be included here […] Southern African Development Community ("SADC") Protocol Against Corruption (2001) and the lesser-known, not in force and barely ratified Economic Community of West African States ("ECOWAS") Protocol on the Fight Against Corruption" (Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 358-359).

[23] "EU Convention Against Corruption Involving EU Officials, the Council of Europe Criminal Law Convention on Corruption, the Council of Europe Civil Law Convention on Corruption and its recommendations." (Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 358).

[24] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 377.

[25] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022), at 382.

[26] Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[27] The authors divided the European instruments into those produced by the European Union (*Convention on the Protection of the European Communities' Financial Interests, Protocol to the Convention on the Protection of the European Communities' Financial Interests, Convention on the Fight Against Corruption involving Officials of the European Communities or Officials of Member States of the European Union, Council Framework Decision 200315681JHA on Combating Corruption in the Private Sector*) and those produced by the Council of Europe (*Twenty Guiding Principles for the Fight Against Corruption, Criminal Law Convention* and *Additional Protocol to the Criminal Law Convention on Corruption, Civil Law Convention on Corruption*), *see* Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[28] The authors listed as part of the OECD instruments the *Convention on Combating Bribery of Foreign Public Officials in International Business Transactions* and added, "The OECD adopted a number of recommendations, such as the 1996 recommendation of the Development Assistance Committee on *Anti-Corruption Proposals for Bilateral Aid Procurement*; the 1998 recommendation on improving ethical conduct in the public service; the 2006 *OECD Council Recommendation on Bribery*

and Africa,[29] as well as the OAS Convention. The second category covers "anti-corruption initiatives in international financial institutions," divided into *World Bank* and "other multilateral financial institutions" initiatives.[30] The last category refers to "private initiatives," encompassing efforts from TI, the ICC, and "other fora," which mention the actions by the Freedom House and the Partnering Against Corruption Initiative (PACI), an initiative from the World Economic Forum (WEF).

Cecily Rose, in turn, conducted an in-depth analysis of the instruments produced by key organizations in developing international anti-corruption law.[31] She focuses on four actors: the OECD, the UN – specifically the UN Office on Drugs and Crime (UNODC), the EITI, and the FATF. Rose highlights the comparisons and contrasts among these international institutions that have approached the creation of relevant anti-corruption norms in distinct ways and at contrasting times. She concludes that the international anti-corruption instruments formulated by these actors, including the non-binding ones, were designed to have domestic consequences, aiming to permeate domestic legal systems by implementing national laws and other regulations prohibiting corrupt conduct. Consequently, these actors hold considerable power as they exercise significant control over the generation of anti-corruption norms that influence domestic legal systems. For Rose, this power, especially concerning domestic legal systems, raises questions about the legitimacy of these institutions and the instruments they produce.

Although several scholars have examined the evolution of the IACR, as described above, there remains a lack of consensus regarding its constituent instruments.[32] In this article, I chose to investigate the incentives for compliance programs in all actors listed by Rose, Villarino, and WRC. While this paper explores

---

and Officially Supported Export Credits; the 2009 *Recommendation of the Council on Tax Measures for Further Combating Bribery of Foreign Public Officials in International Business Transactions*; and the 2009 *Recommendation of the Council for Further Combating Bribery of Foreign Public Officials in International Business Transactions*. In 2010, the OECD adopted the 10 *Principles for Transparency and Integrity in Lobbying*. Corruption is also mentioned in s VII of the *OECD Guidelines for Multinational Enterprises*, which were first adopted in 1976 and updated, for the fifth time, in May 2011. In addition, as was the case for the CoE [Council of Europe], the OECD has published a number of guidelines and tools related to anti-corruption efforts, such as the 'OECD Bribery Awareness Handbook for Tax Examiners' and the *Principles for Donor Action in Anti-Corruption*." (Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013), at 228).

[29] Wouters, Ryngaert, and Cloots listed as part of the African instruments the *African Union Convention on Preventing and Combating Corruption,* the *Southern African Development Community Protocol Against Corruption,* and the *Economic Community of West African States Protocol on the Fight Against Corruption.* (Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013)).

[30] The authors did not detail what these institutions would be, only affirming that "With the World Bank taking on the pioneering role, the other multilateral financial institutions followed suit. The policies of each of the individual institutions cannot be described in detail. Suffice it to say that all these institutions have in some way addressed the problem of corruption. Adopting policies for both internal and/or external corrupt practices." (Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013), at 234).

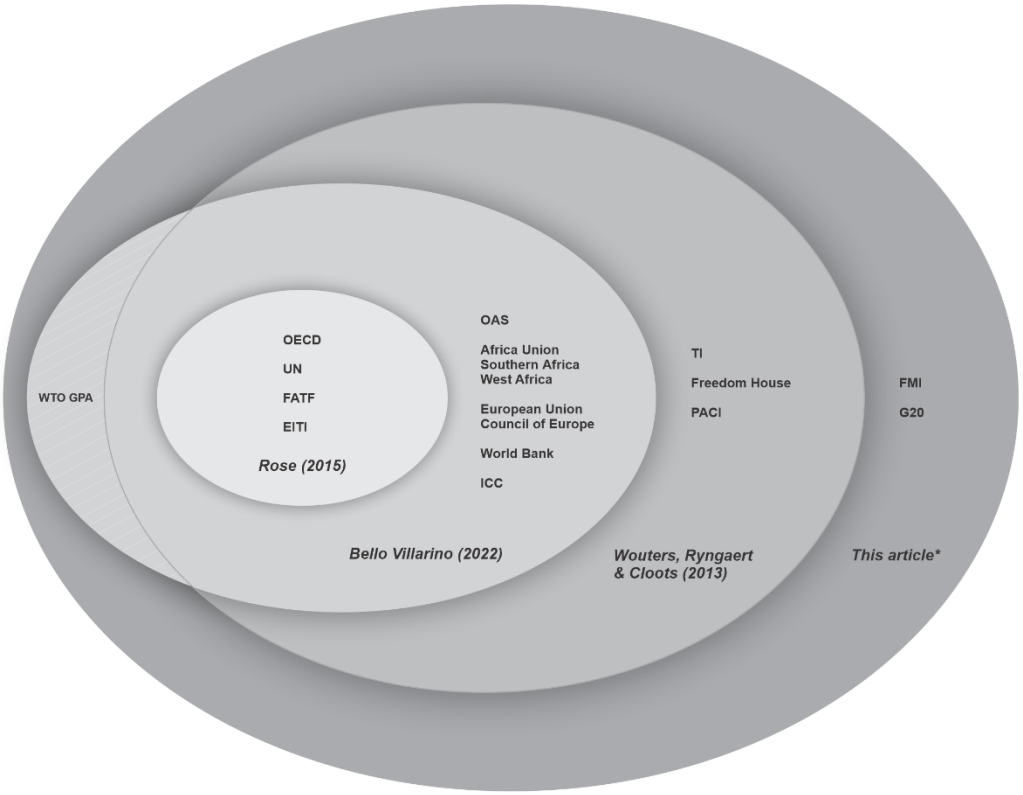
[31] CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015).

[32] The literature analyzed, as well as this article, is not comprehensive in terms of international anti-corruption law as it does not discuss, for instance, some regional conventions such as the *Beijing Declaration on Fighting Corruption* and the *Arab Anti-Corruption Convention*.

the regime listed by these scholars, it goes beyond it by including documents not examined, such as some recommendations, best practices, and guidelines. I examined not only the current versions of these documents but also previous versions to determine *if* and *when* the reference to compliance programs first appeared. I chose to analyze also non-binding documents because, as highlighted by Rose, they have updated and promoted the implementation of conventions, as in the case of the OECD Convention, as well as played a relevant role in shaping anti-corruption domestic laws.[33] Furthermore, I have included two additional actors in this study that the selected literature has not addressed: the Group of Twenty (G20), which has notably promoted compliance programs, and the International Monetary Fund (IMF), which lacks a policy to explicitly promote compliance programs in contrast to the World Bank. Group Figure 1 below illustrates the scope of this article.

[33] *See, e.g.*, KEVIN E. DAVIS, BETWEEN IMPUNITY AND IMPERIALISM: THE REGULATION OF TRANSNATIONAL BRIBERY (2019).

Figure 1: Actors analyzed in the selected literature on the international anti-corruption regime and in this article



I categorized these IACR actors into four groups: international organizations, intergovernmental initiatives, international financial institutions, and international private initiatives. The European Union, which has no mention of compliance programs in its framework, was not included in these categories due to its distinct

characteristics.[34] In the course of this article,[35] among the international organizations analyzed, I could not find incentives for compliance programs in the African framework[36] (Africa Union,[37] Southern Africa,[38] and West Africa[39]) and Council of

[34] On the official European Commission page, the anti-corruption legislation and policies of the European Union are listed in categories (European Commission. *EU Legislation on Anti-corruption.* https://home-affairs.ec.europa.eu/policies/internal-security/corruption/eu-legislation-anti-corruption_en, on 6 Dec. 2023). As main anti-corruption legislations, the following are highlighted: (i) Convention on Fighting Corruption Involving Officials of the EU or Officials of EU Countries (1997); (ii) Council Framework Decision on Combating Corruption in the Private Sector (2003); and (iii) Council Decision 2008/852/JHA. Among the legislation to protect the EU's financial interests: (iv) Directive (EU) 2017/1371; (v) Regulation (EU, Euratom) 2020/2092; (vi) Council Regulation (EU) 2017/1939; (vii) Regulation (EU, Euratom) 883/2013. Within sectoral legislation: (viii) 5th Anti-Money Laundering Directive (AMLD); (ix) Directive (EU) 2018/1673; (x) Directive 2014/42/EU; (xi) Council Decision 2007/845/JHA; (xii) Council Decision 2005/212/JHA; (xii) Regulation (EU) 2018/1805; (xiii) Directive (EU) 2019/1937; EU Rules on Public Procurement (the link directs to the page: European Commission. *Legal Rules and Implementation.* https://single-market-economy.ec.europa.eu/single-market/public-procurement/legal-rules-and-implementation_en, on 6 Dec. 2023, which listed): (xiv) Directive 2014/24/EU, (xv) Directive 2014/25/EU, (xvi) Directive 2014/23/EU; (xvii) Directive (EU) 2010/24; (xiii) Directive (EU) 2011/16. Such legislation does not address anti-corruption compliance programs. The 5th Anti-Money Laundering Directive (AMLD) mentions "the development of internal policies, controls and procedures, including model risk management practices, customer due diligence, reporting, record-keeping, internal control, compliance management including, where appropriate with regard to the size and nature of the business, the appointment of a compliance officer at management level, and employee screening" for the purposes of money laundering and terrorism, including references to FATF Recommendations, which are also mapped in this article. However, as the compliance program mentioned in these documents does not have an anti-corruption purpose, such provisions were not considered for the purposes of this article. On the same official page, Internal Rules for EU Institutions are listed, but they were not analyzed as they are not directed towards corporations. The page also informs that on May 3, 2023, the Commission presented a new Proposal to combat corruption. The Proposal includes provisions for compliance programs. Article 18 stipulates that mitigating circumstances will be considered "where the offender is a legal person and it has implemented effective internal controls, ethics awareness, and compliance programs to prevent corruption prior to or after the commission of the offense" (EUR-Lex. *Document 52023PC0234.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2023%3A234%3AFIN, on 6 Dec. 2023).

[35] For details on how I conducted the search for documents related to compliance programs in each framework, *see* the corresponding footnotes provided for each framework.

[36] The African Framework, as outlined by Villarino and WRC, comprises three documents: (1) ECOWAS Protocol on the Fight Against Corruption (2001); (2) SADC Protocol Against Corruption (2005); (3) AU Convention on Preventing and Combating Corruption (2006).

[37] The Africa Union (AU) Convention on Preventing and Combating Corruption was adopted in 2003 and entered into force in 2006 (TI. (2006). *Anti-corruption Conventions in Africa: What Civil Society Can Do to Make Them Work*. https://uncaccoalition.org/resources/advocacy/anti-corruption-conventions-in-africa.pd). The Convention declares its aim to prevent, detect, punish, and eradicate corruption in Africa's public and private sectors, promote cooperation among state parties, coordinate policies and legislation, and foster socio-economic development while ensuring transparency and accountability in public affairs (AU. (2003). *African Union Convention on Preventing and Combating Corruption*. https://au.int/sites/default/files/treaties/36382-treaty-0028_-_african_union_convention_on_preventing_and_combating_corruption_e.pdf). The AU Convention emphasize the need for states to implement measures against corruption in the private sector. However, it does not explicitly refer to compliance programs.

[38] The Southern African Development Community (SADC), founded in 1992, has 26 protocols as part of a legally binding document committing member states to certain objectives and specific procedures, among them the Protocol on the Fight Against Corruption (SADC. SADC Protocols. https://www.sadc.int/pages/sadc-protocols, on 9 Aug. 2023). This document does not mention compliance programs.

[39] The Economic Community of West African States (ECOWAS) Protocol on the Fight Against Corruption was adopted by ECOWAS member states in 2001 but faced challenges in reaching the

Europe framework.[40] Within intergovernmental initiatives, there were no identified incentives for compliance programs in the Revised WTO GTA[41] and the FATF frameworks.[42] Among international financial institutions, I did not find at the IMF

required threshold to come into force (TI. *ECOWAS Protocol on the Fight Against Corruption* (2001). https://knowledgehub.transparency.org/guide/international-anti-corruption-commitments/8008, on 9 Aug. 2023). It does not mention compliance programs.

[40] The Council of Europe is an international organization based in France that protects human rights, democracy and the rule of law (Consilium. *European Council and Council of the European Union: What's the difference?* https://www.consilium.europa.eu/en/european-council-and-council-of-the-eu/#:~:text=Council%20of%20the%20European%20Union%20('the%20Council')&text=The%20Council%20of%20Europe%20is,and%20the%20rule%20of%20law, on 6 Dec. 2023). According to WRC, the Council of Europe framework is composed of the following documents: (i) Criminal Law Convention and Additional Protocol to the Criminal Law Convention on Corruption; (ii) Civil Law Convention on Corruption; (iii) Twenty Guiding Principles for the Fight Against Corruption. No one mention compliance programs. One of the monitoring bodies of the Council of Europe is the Group of States Against Corruption (GRECO). In addition to the documents cited by WRC, they included, among the anti-corruption legal instruments adopted by the Council of Europe, (iv) the Recommendation on Codes of Conduct for Public Officials and (v) the Recommendation on Common Rules Against Corruption in the Funding of Political Parties and Electoral Campaigns (Council of Europe. *Group of States Against Corruption*. https://www.coe.int/en/web/greco/home, on 6 Dec. 2023). Both do not mention compliance programs either.

[41] Created in 1995, the WTO provides a forum for negotiating agreements aimed at reducing obstacles to international trade and ensuring a level playing field, thus contributing to economic growth and development (WTO. Overview. https://www.wto.org/english/thewto_e/whatis_e/wto_dg_stat_e.htm, on 9 Aug. 2023). The WTO also establishes the legal and organizational structure to implement, monitor, and resolve conflicts related to these agreements. Presently, the WTO's collection of trade pacts includes 16 multilateral agreements (applicable to all WTO members) and two plurilateral agreements (to which only some WTO members are parties) (WTO. Overview. https://www.wto.org/english/thewto_e/whatis_e/wto_dg_stat_e.htm, on 9 Aug. 2023). Villarino includes one of these plurilateral agreements, the Revised WTO Global Procurement Agreement (WTO Revised GPA), within the IACR. The primary objective of this agreement is to facilitate the reciprocal opening of government procurement markets among its participating parties (WTO. *Agreement on Government Procurement*. https://www.wto.org/english/tratop_e/gproc_e/gp_gpa_e.htm, on 9 Aug. 2023). The WTO Revised GPA was published in 2012 and entering into force in 2014 (WTO. *Revised Agreement on Government Procurement*. https://www.wto.org/english/tratop_e/gproc_e/memobs_e.htm, on 9 Aug. 2023). In the preamble, the WTO Revised GPA recognizes the importance of transparent measures regarding government procurement, conducting procurements in a transparent and impartial manner, and avoiding conflicts of interest and corrupt practices in accordance with applicable international instruments, such as the UN Convention. Moreover, one of the general principles of the protocol is the conduct of procurement, which include that a procuring entity shall conduct it in a transparent and impartial manner, preventing corrupt practices. However, there is no specific mention regarding corporate compliance programs.

[42] FAFT, established in 1989, is an intergovernmental body that defines itself as "the global money laundering and terrorist financing watchdog." (FATF. *Our Topics*. https://www.fatf-gafi.org/en/home.html, on 20 Aug. 2023). Villarino, Rose, and WRC include FATF at the IACR. FATF does not directly address corruption. However, the organization understands that corruption and money laundering are often intrinsically linked, as corruption-related offenses are generally committed with the purpose of gaining illicit funds, and money laundering is the procedure used to conceal the origin of those funds obtained through illegal activity (FATF. *Corruption*. https://www.fatf-gafi.org/en/topics/corruption.html, on 20 Jun. 2023). FATF's objectives were outlined in the document 40 FATF Recommendations, which aims to "provide a comprehensive framework of measures to help countries tackle illicit financial flows" (FATF. *The FATF Recommendations.* https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html, on 20 Jun. 2023). The Recommendations has four versions, from 1990, 1996, 2003 and 2012 (FATF. *Review of the FATF Standards and Historical Versions.* https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Fatfrecommendations/Review-and-history-of-fatf standards.html, on 20 Jun. 2023). FATF emphasizes that the 40 Recommendations play a crucial role in combating corruption by promoting transparency within the financial system, enabling easier detection, investigation, and

specific documents regarding policies to promote compliance programs.[43] Concerning international private initiatives, I found no documents directly addressing compliance

prosecution of corruption connected to money laundering cases, and facilitating the recovery of illicitly acquired assets (FATF. *Corruption*. https://www.fatf-gafi.org/en/topics/corruption.html, on 20 Jun. 2023). Furthermore, FATF affirms the interrelationship between the IACR and its own framework, affirming, for instance, that the implementation of the UN Convention, including the non-binding provisions such as the establishment of, in member countries, financial intelligence units responsible for receiving, analyzing and disseminating reports of suspicious financial transactions to the competent authorities, would complement a jurisdiction's anti-money laundering program (OECD/FATF. (2012). *FATF Report: Specific Risk Factors in Laundering the Proceeds of Corruption Assistance to Reporting Institutions*. https://www.fatf-gafi.org/en/publications/Methodsandtrends/Specificriskfactorsinthelaunderingofproceedsofcorruption-assistancetoreportinginstitutions.html, on 20 Jun. 2023). FATF, in addition to setting international standards, conducting evaluations, and promoting measures to combat money laundering and terrorist financing on a global scale, also establishes guidelines and best practices. Among the 37 guidelines and best practices produced (FATF. *The FATF Recommendations – Guidance and Best Practices.* https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html, on 20 Jun. 2023), only one deals specifically with corruption: "Best Practices Paper: The Use of the FATF Recommendations to Combat Corruption" (FATF. (2013). *Best Practices Paper: The Use of the FATF Recommendations to Combat Corruption*. https://www.fatf-gafi.org/en/publications/Corruption/Bpp-fatfrecs-corruption.html, on 20 Jun. 2023). FATF 40 Recommendations and their associated guidelines and best practices stipulate that certain private sector institutions should establish a compliance program against money laundering, which holds the potential to uncover corruption that involves money laundering. However, FATF framework does not directly promote anti-corruption compliance programs.

[43] The IMF's efforts in combating corruption have been subject to some criticism (*e.g.*, TI, Human Right Watch, and Global Witness. (2020). Letter to IMF Executive Board on April 08, 2020. https://images.transparencycdn.org/images/TI_HRW_GW_Letter_IMF_COVID19_Emergency_Funding.pdf). In fact, the IMF has historically refrained from explicitly using the word "corruption" in its reports and actions, avoiding dealing with the problem (Ivana M. Rossi & Jonathan Pampolina. (2023). *Taking Stock of the Governance and Anti-Corruption Work of the IMF and the Way Forward [Event].* OECD. https://www.oecd-events.org/gacif2023/session/46641e7d-1af6-ed11-907a-000d3a474dec, on 11 Aug. 2023). The shift towards a stronger stance against corruption began in 2018 when the IMF Executive Board adopted the Framework for Enhanced Engagement on Governance "to promote more systematic, effective, candid, and evenhanded engagement with member countries regarding corruption of macro critical dimensions and governance vulnerabilities linked to corruption" (IMF. (2023). *Press Release No. 23/115.* https://www.imf.org/en/News/Articles/2023/04/11/pr23115-imf-executive-board-concludes-review-implementation-framework-enhanced-engagement-governance#:~:text=In%20April%202018%2C%20the%20IMF,governance%20vulnerabilities%20linked%20to%20corruption, on 11 Aug. 2023). The 2018 Framework includes a systematic assessment of governance vulnerabilities and corruption with respect to all members, as well as an assessment of governmental measures to prevent bribery. Nevertheless, the 2018 framework was implemented weakly, partly due to the pandemic (Ivana M. Rossi & Jonathan Pampolina. (2023). *Taking Stock of the Governance and Anti-Corruption Work of the IMF and the Way Forward [Event].* OECD. https://www.oecd-events.org/gacif2023/session/46641e7d-1af6-ed11-907a-000d3a474dec, on 11 Aug. 2023). In 2023, the IMF published the Review of the Implementation of the 2018 Framework (IMF. (2023). *Review of Implementation of the 2018 Framework for Enhanced Fund Engagement on Governance*. https://www.imf.org/en/Publications/Policy-Papers/Issues/2023/04/11/Review-of-Implementation-of-The-2018-Framework-for-Enhanced-Fund-Engagement-on-Governance-532166, on 11 Aug. 2023). Under these documents, there is no direct incentive for corporate compliance programs.

programs in the EITI[44] and the Freedom House[45] frameworks. In the subsequent section, I will analyze the documents within IACR where I could find direct incentives for compliance programs.

II. MAPPING COMPLIANCE PROGRAM IN THE IACR

In this section, I present the mapping of documents produced by the selected actors, as described in Section 2, aiming to gain insights into their treatment of corporate compliance programs. While the focus is on documents that expressly mention compliance programs, I included some documents that do not cite this strategy to provide a historical perspective on the actor's approach to the treatment of this strategy. The presentation of the documents by each actor follows the chronological order of publication of the analyzed documents. The Appendix to this article provides a summary of the documents described in this section that promote compliance programs.

---

[44] EITI is a multi-stakeholder organization registered as a non-profit association in 2003, (EITI. *Governance.* https://eiti.org/governance, on 3 Jul. 2023) which aims to enhance transparency and accountability in the extractive sector (oil, gas, and mineral industries) (EITI. Our mission. https://eiti.org/our-mission, on 3 Jul. 2023). The EITI Standard "requires the disclosure of information along the extractive industry value chain from the point of extraction, to how revenues make their way through the government, and how they benefit the public" (DoJ. *Extractive Industries Transparency Initiative (EITI)*. https://www.state.gov/extractive-industries-transparency-initiative-eiti/#:~:text=The%20EITI%20Standard%20requires%20the,how%20they%20benefit%20the%20public, on 3 Jul. 2023). The EITI Standard underwent four revisions, 2013, 2016, 2019, and 2023 (according to a search on the EITI website for the expression "EITI Standard," on the field "Search," on 3 Jul. 2023, *see* https://eiti.org/search?content-type[document]=document&viewsreference[enabled_settings][argument]=argument). The 2023 version is the one that cited corruption more times, stating expectation that companies that support the EITI Standard "engage in rigorous due diligence processes and publish an anti-corruption policy setting out how the company manages corruption risk, including how the company collects and takes risk-based steps to use beneficial ownership data regarding joint venture partners, contractors, and suppliers in its processes" (EITI. (2023). *The EITI Standard 2023* – Part 1. https://eiti.org/sites/default/files/2023-06/2023%20EITI%20Standard.pdf, at 39). While the current version highlights the contribution of the EITI Standard in the fight against corruption and encourages companies to adopt anti-corruption policies, it does not include specific provisions about compliance programs.

[45] The Freedom House is a non-governmental organization (NGO) and research institute based in the United States, which WRC includes within the IACR. Established in 1941, it has become a prominent American organization dedicated to advocating, developing programs, and conducting research in support of democracy worldwide (Freedom House. *About us*. https://freedomhouse.org/about-us, on 16 Oct. 2023). One of the policy recommendations from Freedom House centers on combating corruption (Freedom House. Policy Recommendations: Combatting Corruption. https://freedomhouse.org/policy-recommendations/combatting-corruption-and-kleptocracy, on 16 Oct. 2023). The Freedom House understands that the corruption and kleptocracy can posing a significant threat to democracy worldwide, as "corruption undermines the freedom and the interests of ordinary citizens, and the effects are especially harmful in developing countries with limited resources and weaker anticorruption mechanisms." While the Freedom House Policy Recommendation on Combatting Corruption provides various suggestions for U.S. policymakers concerning anti-corruption laws, there is no specific mention of corporate compliance programs.

A. International Organizations

1. OAS

The OAS is the world's oldest regional organization globally.[46] Currently, the OAS unites all 35 independent states of the Americas, serving as the region's primary political, juridical, and social governmental forum. [47]

a. OAS Convention (1996)

The OAS Convention – adopted during the General Assembly of the OAS held in Venezuela in 1996 and coming into force in 1997[48] – is considered the oldest legal instrument in the IACR.[49] Its preamble declares that the member states of OAS are "convinced that corruption undermines the legitimacy of public institutions and strikes at society, moral order and justice, as well as at the comprehensive development of peoples."[50] The OAS Convention aims to address it to strengthen mechanisms to prevent, detect, punish, and eradicate corruption while fostering cooperation among member states to combat corruption in public functions and related acts.[51] Currently, all countries in the Americas are signatories of the OAS Convention, except Cuba.[52]

The OAS Convention consists of provisions that impose varying degrees of obligation on its members.[53] Among the provisions, the states must reform their legal system to make illegal offenses involving both active and passive bribery of public officials, whether domestic or foreign, including holding companies liable for these

---

[46] OAS dating back to the First International Conference of American States, held in Washington, D.C., from October 1889 to April 1890. Its primary objectives are to foster peace and justice among its member states, encourage solidarity and enhance collaboration with Sovereignty. (OAS. *Who we are?* https://www.oas.org/en/about/who_we_are.asp, on 2 Jun. 2023).

[47] OAS. *Who we are?* https://www.oas.org/en/about/who_we_are.asp, on 2 Jun. 2023. OAS countries members: Antigua and Barbuda, Argentina, Bahamas (Commonwealth of), Barbados, Belize, Bolivia, Brazil, Canada, Chile, Colombia, Costa Rica, Cuba, Dominica (Commonwealth of), Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Guyana, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Saint Lucia, San Kitts and Nevis, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, United States of America, Uruguay, Venezuela (Bolivarian Republic of). (OAS. *Member States*. https://www.oas.org/en/member_states/default.asp, on 2 Jun. 2023.)

[48] OAS. *Background*. https://www.oas.org/juridico/english/corr_bg.htm, on 2 Jun. 2023.

[49] Altamirano, G. D. (2006). The Impact of The Inter-American Convention Against Corruption. *University of Miami Inter-American Law Review*, 38(3), 487-548; Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022).

[50] OAS (1996). *Inter-American Convention Against Corruption*. https://www.oas.org/en/sla/dil/docs/inter_american_treaties_B-58_against_Corruption.pdf.

[51] Article II. OAS (1996). *Inter-American Convention Against Corruption*. https://www.oas.org/en/sla/dil/docs/inter_american_treaties_B-58_against_Corruption.pdf.

[52] OAS. *B-58 Signatories and Ratifications.* https://www.oas.org/en/sla/dil/inter_american_treaties_B-58_against_Corruption_signatories.asp, on 27 Jun. 2023.

[53] Giorleny D. Altamirano, *The Impact of the Inter-American Convention Against Corruption*, 38 UNIVERSITY OF MIAMI INTER-AMERICAN LAW REVIEW 487 (2006); Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022).

illicit activities.[54] The OAS Convention does not mention compliance programs.

b. OAS Convention Recommendations (2004, 2006, 2010, 2015)

In 2001, the OAS General Assembly adopted the Report of Buenos Aires on the Mechanism for Follow-up on Implementation of the Inter-American Convention Against Corruption, which defines the structure and elements of the Follow-up Mechanism for the Inter-American Convention Against Corruption (MESICIC).[55] The follow-up mechanism operates through voluntary peer reviews, examining member countries' domestic laws and institutions to determine if they accord with the provisions of the Convention.[56] The OAS follow-up mechanism involves monitoring and assisting national governments in implementing the OAS Convention and harmonizing anti-corruption regulations across member states.[57]

The MESICIC issued four recommendations (2004, 2006, 2010, 2015), aiming to enhance and align with the provisions of the OAS Convention.[58] Although compliance programs are not explicitly mentioned in the recommendations, the last two editions emphasize the importance of private sector involvement in combating corruption to achieve the OAS Convention's objectives.[59] In addition, in this recommendations, both the Committee and member states are urged to promote anti-corruption practices within the private sector, especially those that identify internal corruption and enable reporting of misconduct to relevant authorities.

c. Other OAS MESICIC Documents

The MESICIC recognizes best practices in anti-corruption public policies within member states, including the promotion of compliance programs.[60] It

---

[54] Article IV and VIII. OAS (1996). *Inter-American Convention Against Corruption*. https://www.oas.org/en/sla/dil/docs/inter_american_treaties_B-58_against_Corruption.pdf.

[55] MESICIC is the acronym for *Mecanismo de Seguimiento de la Implementación de la Convención Interamericana contra la Corrupción*, the name of the OAS follow-up mechanism in Spanish.

[56] Giorleny D. Altamirano, *The Impact of the Inter-American Convention Against Corruption*, 38 UNIVERSITY OF MIAMI INTER-AMERICAN LAW REVIEW 487 (2006).

[57] Giorleny D. Altamirano, *The Impact of the Inter-American Convention Against Corruption*, 38 UNIVERSITY OF MIAMI INTER-AMERICAN LAW REVIEW 487 (2006).

[58] The Recommendations are drafted at the Meeting of the Conference of the States Parties under MESICIC, which has had four editions: (1) First Meeting: Washington D.C. April 1 - 2, 2004; (2) Second Meeting: Washington D.C. November 20 - 21, 2006; (3) Third Meeting: Brasilia, D.F. December 9 - 10, 2010; (4) Fourth Meeting: Washington, D.C. December 14 - 15, 2015. Two documents were produced during each meeting: one containing the recommendations and another with the final minutes. I analyzed all eight documents. (OAS. *Anticorruption Portal of the Americas – The MESICIC in Documents – Meetings and recommendations.* http://www.oas.org/en/sla/dlc/mesicic/documentos_recomendaciones.html, on 19 July 2023).

[59] OAS. (2010). *Recommendations of the Third Meeting of the Conference of States Parties of the MESICIC*. http://www.oas.org/en/sla/dlc/mesicic/docs/cepIII_recom_en.pdf; OAS. (2015). *Recommendations of the Fourth Meeting of the Conference of States Parties of the MESICIC*. http://www.oas.org/en/sla/dlc/mesicic/docs/mesicic_cosp_iv_rec_eng.pdf.

[60] MESICIC also recognizes the state members' best practices and classifies them into 17 subjects. Topics on best practices: (1) Government hiring; (2) Standards of conduct to prevent conflict of interest in the public administration; (3) Understanding of ethical rules and responsibilities by public servants; (4) Equitable remuneration and probity in public service; (5) Disclosure of income, assets, and liabilities by persons who perform public functions; (6) Official duty to report acts of corruption; (7) Protection of those who report acts of corruption; (8) Rules for the conservation of public resources; (9) Government procurement; (10) Denial of favorable tax treatment for expenditures made in violation of

recognized the Mexican best practice named Register of Business Integrity, which encourages companies to combat corruption, implement ethics and integrity codes, and implement best practices.[61] The MESICIC also presents the Paraguayan initiative "*Sello Integridad*." [62] It is a biennial recognition awarded by the Paraguayan anti-corruption agency (*Secretaría Nacional Anticorrupción*) to companies that have compliance programs aimed to promote businesses to adopt measures and actions to prevent, detect, and remedy acts of corruption and fraud, as well as efforts to foster an organizational culture of integrity.[63]

2. OECD

The OECD is an international organization that aims to improve economic

---

anticorruption laws; (11) Prevention of bribery of domestic and foreign government officials; (12) Mechanisms to encourage participation by civil society and nongovernmental organizations in efforts to prevent corruption; (13) Criminalization of acts of corruption; (14) Criminalization of transnational bribery; (15) Criminalization of illicit enrichment; (16) Mutual technical cooperation and reciprocal assistance; (17) Extradition of those who commit acts of corruption. (OAS. *Anticorruption Portal of the Americas – Best Practices to Prevent and Combat Corruption.* http://www.oas.org/en/sla/dlc/mesicic/buenas-practicas.html, on 19 Jul. 2023). On July 19, 2023, given the large number of best practices documents presented by MESICIC, I applied certain criteria to select which ones to analyze. Firstly, I examined the subjects of the best practices and selected subjects 7, 9, 10, 11, and 12 (*see supra* note), as the others were related to public functions or acts that did not seem to encompass any relation with the private sector. Concerning these selected subjects, I read the titles of all the best practices and selected the ones related to compliance programs or the private sector. Regarding the 7th subject, "protection of those who report acts of corruption," I selected and read the full documents of the following best practices: "*Paraguay: Portal de Denuncias Anticorrupción y Sistema Informático de Registro y Seguimiento de Causas*," "*Mexico: System of Internal and External Whistleblowers of Corruption*," and "*Panamá: Recepción de Denuncias mediante Plataforma electrónica Tu Pista Administrada por Crime Stoppers*." Only the Paraguayan document mentioned compliance programs, but it concerned another best practice, the "*Sello Integridad*," which is part of the 12th subject, as I will explain. Regarding the 9th subject, "government procurement," I selected and attempted to read "*República Dominicana: Experiencia de uso en ciencias de los datos y compliance para la prevención de la* [incomplete]," but the two related documents were not available. As for the 10th subject, "denial of favorable tax treatment for expenditures made in violation of anticorruption laws," no best practice was listed. Regarding the 11th subject, "prevention of bribery of domestic and foreign government officials," I read "Mexico: International Certification ISO 37001: 2016 'Anti-bribery Management Systems' of the General [incomplete]," which is about the certification of a public agency and was not pertinent to this search. Finally, concerning the 12th subject, "Mechanisms to encourage participation by civil society and nongovernmental organizations in efforts to prevent corruption," I read the full text of the best practices: "*Paraguay: Sello Integridad*" and "Mexico: Register of Business Integrity." Both practices were about corporate compliance programs and I have described them above.

[61] OAS. (2019). *Mexico: Register of Business Integrity – Best Practices Form: Learn about the objectives and results of this strategy.* http://www.oas.org/en/sla/dlc/mesicic/docs/bp2019_sem2_mex_form_mecparticip2.pdf; OAS. (2019). *Mexico: Register of Business Integrity – Annex to the form: Presentation.* http://www.oas.org/en/sla/dlc/mesicic/docs/bp2019_sem2_mex_ppt2.pdf.

[62] OAS. (2023). *Paraguay: Sello Integridad – Best Practices Form: Learn about the Objectives and Results of this Strategy.* http://www.oas.org/en/sla/dlc/mesicic/docs/mar2023_bp_py%20_form.pdf; OAS. (2023). *Paraguay: Sello Integridad – Presentation.* http://www.oas.org/en/sla/dlc/mesicic/docs/mar2023_bp_py%20_ppt.pdf.

[63] MESICIC also has two draft regulations, one on the declaration of income, assets, and liabilities and another on the protection of whistleblowers; both aim to reflect the highest international standards and serve as models that OAS state members can utilize when drafting anti-corruption laws. The first one is targeted at those who perform public duties. The second one addresses an aspirational provision of the OAS Conviction, which states that parties should consider measures within their institutional systems

performance in the world, providing a forum and knowledge hub for data and analysis, exchange of experiences, best-practice sharing, and advice on public policies and international standard-setting.[64] Currently, the OECD has 38 countries members.[65] The OECD has produced around 460 legal instruments in several subjects.[66] International agreements and decisions are legally binding for OCDE parties' members and other countries that adhere to it.[67] Recommendations and guidelines are not legally binding, but OECD membership implies an expectation that member states will do their best to implement them. Concerning the IACR, the OECD anti-corruption efforts are significant for two main reasons.[68] First, as indicated above, the OECD was a driving force in the expansion of anti-bribery law. Secondly, several of the most prominent players in international trade are OECD member states.[69]

a. OECD Convention (1997)

The OECD Convention, adopted in 1997 and enforced in 1999, emerged during an era when bribery was widely accepted and even tax-deductible in certain countries.[70] The Convention's main objective was to level the playing field among a limited number of countries "in a position to regulate the supply of bribes by corporations involved in international business transactions."[71] The Convention, compulsory for signatories, mandates that states criminalize the act of bribing foreign public officials in

---

to create, maintain, and strengthen protection systems for those who report acts of corruption (Article II, 8). It also concerns the mechanisms states should develop within their structures to enhance the fight against corruption. MESICIC also provides legislative guidelines with elements that states should consider when formulating laws related to the matters addressed in the OAS Convention. All these legislative guidelines are about creating obligations for the public administration and do not mention compliance programs to companies. (OAS. *Anticorruption Portal of the Americas – Legal Cooperation Tools – Legislative Guidelines.* http://www.oas.org/en/sla/dlc/mesicic/leyes.html, on 19 Jul. 2023).

[64] OECD. *Who we are.* https://www.oecd.org/about/members-and-partners/, on 27 Jul. 2023.

[65] OECD members: Australia, Austria, Belgium, Canada, Chile, Colombia, Costa Rica, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Türkiye, United Kingdom, and the United States. (OECD. *About*. https://www.oecd.org/about/, on 22 Jun. 2023).

[66] OECD. *OECD Legal Instruments*. https://legalinstruments.oecd.org/en/about, on 3 Feb. 2022.

[67] If a member chooses to abstain during the decision-making process, it will not be legally binding for them. (OECD. *OECD Legal Instruments*. https://legalinstruments.oecd.org/en/about, on 3 Feb. 2022.). The 23-decision published by OECD does not cite anti-corruption corporate compliance programs or functional equivalents. I searched it on June 29, 2023, on the *Compendium of OECD Legal Instruments*, available at https://legalinstruments.oecd.org/en/, on 3 Feb. 2022, for the word "corruption" and also "compliance," one at a time, using the filter "decision" on the field "type(s)."

[68] Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[69] Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[70] Humboldt-Viadrina School of Governance. (2013). *Motivating Business to Counter Corruption: A Practioner Handbook on Anti-Corruption Incentives and Sanctions*. https://www.globalcompact.de/migrated_files/wAssets/docs/Korruptionspraevention/Publikationen/motivating_business_to_counter_corruption.pdf; Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022).

[71] CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015), at 60.

international business transactions.[72] It does not mention compliance programs.

Following the Convention, the OECD published several recommendations and guidance documents aimed at furthering the development of anti-corruption standards among state parties.[73] These documents will be presented chronologically bellow.[74] The OECD Convention follow-up mechanism, the Working Group on Bribery in

---

[72] OECD. (1997). *Convention on Combating Bribery of Foreign Public Officials in International Business Transactions.* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0293*,*

[73] CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015).

[74] I searched for non-binding documents referencing corporate anti-corruption compliance programs on March 8, 2022. I did this search manually on the OECD website in an exploratory manner, and I not intended to be exhaustive. I excluded documents from the search result that did not discuss anti-corruption compliance (such as those related to tax compliance) or recommended anti-corruption compliance for organizations other than those targeted in this article (such as state-owned enterprises, public entities, and development agencies). On June 29, 2023, I performed another search to verify and update the previously obtained results. This time, I utilized the Compendium of OECD Legal Instruments, which was accessible at https://legalinstruments.oecd.org/en/, on 8 Mar. 2022. I searched for "corruption" and "compliance" without applying any filters. The search yielded 28 results, namely(1) OECD/LEGAL/0378Recommendation of the Council for Further Combating Bribery of Foreign Public Officials in International Business Transactions; (2) OECD/LEGAL/0144 Declaration on International Investment and Multinational Enterprises; (3) OECD/LEGAL/0421Declaration on the Fight Against Foreign Bribery - Towards a New Era of Enforcement; (4) OECD/LEGAL/0451Recommendation of the Council on Guidelines on Anti-Corruption and Integrity in State-Owned Enterprises; (5) OECD/LEGAL/0431Recommendation of the Council for Development Co-operation Actors on Managing the Risk of Corruption; (6) OECD/LEGAL/0447 Recommendation of the Council on Bribery and Officially Supported Export Credits; (7) OECD/LEGAL/0413 Recommendation of the Council on Principles of Corporate Governance; (8) OECD/LEGAL/0316 Recommendation of the Council on OECD Guidelines for Managing Conflict of Interest in the Public Service; (9) OECD/LEGAL/0414 Recommendation of the Council on Guidelines on Corporate Governance of State-Owned Enterprises; (10) OECD/LEGAL/0349 Recommendation of the Council on Principles for Private Sector Participation in Infrastructure; (11) OECD/LEGAL/0411Recommendation of the Council on Public Procurement; (12) OECD/LEGAL/0392 Recommendation of the Council on Principles for Public Governance of Public-Private Partnerships; (13) OECD/LEGAL/0435 Recommendation of the Council on Public Integrity; (14) OECD/LEGAL/0298 Recommendation of the Council on Improving Ethical Conduct in the Public Service Including Principles for Managing Ethics in the Public Service; (15) OECD/LEGAL/0396 Recommendation of the Council on Fighting Bid Rigging in Public Procurement; (16) OECD/LEGAL/0476 Recommendation of the Council on Foreign Direct Investment Qualities for Sustainable Development; (17) OECD/LEGAL/0379 Recommendation of the Council on Principles for Transparency and Integrity in Lobbying; (18) OECD/LEGAL/0469 Recommendation of the Council on the Ten Global Principles for Fighting Tax Crime; (19) OECD/LEGAL/0486Recommendation on the Role of Government in Promoting Responsible Business Conduct; (20) OECD/LEGAL/0269 Recommendation of the Council concerning an OECD Model Agreement for the Undertaking of Simultaneous Tax Examinations; (21) OECD/LEGAL/0369 Recommendation of the Council on Enhancing Integrity in Public Procurement; (22) OECD/LEGAL/0445 Recommendation of the Council on Public Service Leadership and Capability; (23) OECD/LEGAL/0438 Recommendation of the Council on Open Government; (24) OECD/LEGAL/0452 Recommendation of the Council concerning Effective Action Against Hard Core Cartels; (25) OECD/LEGAL/0327 OECD Principles of Corporate Governance; (26) OECD/LEGAL/0282 Recommendation of the Council for Facilitating International Technology Co-operation with and among Businesses; (27) OECD/LEGAL/0444 Recommendation of the Council on Global Events and Local Development; (28) OECD/LEGAL/0337 Recommendation of the Council on OECD Guidelines on Corporate Governance of State-Owned Enterprises. By applying the same exclusion criteria used in the 2022 manual search, I did not find any additional results beyond those obtained in the initial search. On June 29, 2023, I conducted another search to find updates to the documents I had initially selected in 2022. During this search, I came across the 2023 version of the Declaration on International Investment and Multinational Enterprises and the Revised Recommendation of The Council on Principles of Corporate Governance, both of which were included in the search results.

International Business Transactions (hereinafter Working Group), is responsible for monitoring the implementation and enforcement of the OECD Convention by peer review,[75] as well as push members to incorporated or given legal effect to these subsequent non-binding instruments within their domestic frameworks.[76]

b. OECD Recommendation of the Council for Further Combating Bribery of Foreign Public Officials in International Business Transactions (1997, 2009, 2021)[77]

The first version is from 1997 is named the Revised Recommendation of the Council on Bribery in International Business Transactions.[78] The document recommends that states encourage companies to establish internal control mechanisms, set standards of conduct, and disclose such actions in annual reports. The 1997 Recommendation also suggests that companies be encouraged to have communication channels and protection measures for those who do not wish to violate professional or ethical standards due to orders or pressure from superiors. In other words, while there is no explicit mention of compliance programs (the word "compliance" does not even appear in the document), there is an indication that the state should stimulate companies to adopt anti-bribery mechanisms.

The 2009 Recommendation explicitly mentions "compliance programs," as it recommends that states encourage companies to develop and adopt adequate internal controls, ethics, and compliance programs or measures to prevent and detect bribery of foreign public officials.[79] This recommendation has the annexed Good Practice Guidance on Internal Controls, Ethics, and Compliance. This guide was formulated based on conclusions and recommendations from the Working Group and is directed at companies and professional associations, recognizing the essential role of the private sector in achieving the objectives of the OECD Convention.

In 2016, the Ministers and Representatives of the Parties to the OECD Convention made a Ministerial Declaration, title Fight Against Foreign Bribery:

---

[75] OECD. *OECD Working Group on Bribery in International Business Transactions.* https://www.oecd.org/corruption/anti-bribery/anti-briberyconvention/oecdworkinggrouponbriberyininternationalbusinesstransactions.htm, on 31 Jun. 2023.

[76] CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015).

[77] Three OECD Recommendations complement the OECD Convention: *The Recommendation of The Council for Development Co-Operation Actors on Managing the Risk of Corruption* [OECD/LEGAL/0431], *The Recommendation of the Council on Tax Measures for Further Combating Bribery of Foreign Public Officials in International Business Transactions* [OECD/LEGAL/0371], *The Recommendation of the Council on Bribery and Officially Supported Export Credits* [OECD/LEGAL/0447]. The last two were not described in this article because they do not mention compliance programs.

[78] OECD. (1997) *1997 Revised Recommendation of the Council on Bribery in International Business Transactions*. https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=C(97)123/FINAL&docLanguage=En.

[79] OECD. (2009). *OECD Recommendation for Further Combating Bribery of Foreign Public Officials in International Business Transactions*. https://www.oecd.org/investment/anti-bribery/anti-briberyconvention/oecdantibriberyrecommendation2009.htm#:~:text=About%20the%202009%20Recommendation&text=The%20Recommendation%20was%20adopted%20by,Internal%20Controls%2C%20Ethics%20and%20Compliance.

Towards a New Era of Enforcement.[80] It expressed appreciation for the Working Group's thorough analysis of issues concerning to several topics, including anti-corruption compliance programs, and encouraged the Working Group to further investigate good practices related to these matters. In addition, the Declaration invites the business community to increase cooperation with governments in the battle against foreign bribery and corruption. It also encouraged companies to widely adopt the OECD Good Practice, annexed to the 2009 Recommendation. Moreover, it urged continuous international initiatives aimed at identifying and promoting effective strategies to prevent foreign bribery and corruption. This includes the implementation of anti-corruption compliance measures and codes of conduct, as well as suitable safeguards in public procurement processes.

The 2021 Recommendation, currently in force, expands the guidance for states to promote the adoption of compliance programs.[81] It has a section called "incentives for compliance," wherein it recommends that member countries should encourage their government agencies to consider compliance programs in their decisions to grant public advantages – including subsidies, licenses, contracts, and export credits – especially in the context of international business transactions. Moreover, the document recommends that such anti-corruption mechanisms should also be considered when applying penalties related to the bribery of foreign public officials, including as a potential mitigating factor. Furthermore, affirms that states must provide adequate training for their authorities to consider such mechanisms in decision-making processes, as well as to ensure easily accessible guidance on these benefits for companies. Regarding public procurement, the 2021 Recommendation states that countries should enact legislation allowing authorities to suspend or prevent participation in public procurements of companies that have bribed foreign public officials, considering the compliance programs as mitigating factors for such sanctions. Another innovation of the 2021 Recommendation is the guidance for considering the remediation measures adopted by companies, including compliance programs, in non-trial resolutions with companies that have bribed foreign public officials. This version also includes an annex, similar to the 2009 Recommendation, named Good Practice Guidance on Internal Controls, Ethics, and Compliance.

c. Declaration on International Investment and Multinational Enterprises (1976, amended in 1979, 1984, 1991, 2000, 2011, 2023)

Since its first adoption in 1976, the Declaration has been a commitment from member states "to provide an open and transparent environment for international investment and to encourage the positive contribution multinational enterprises can make to economic and social progress."[82] It was amended in 1979, 1984, 1991, 2000,

---

[80] OECD. (2016). *Ministerial Declaration – The Fight Against Foreign Bribery: Towards a New Era of Enforcement*. https://www.oecd.org/corruption/OECD-Anti-Bribery-Ministerial-Declaration-2016.pdf.

[81] OECD. (2021). *Recommendation of the Council for Further Combating Bribery of Foreign Public Officials in International Business Transactions*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0378, 8 on Mar. 2022.

[82] OECD. *OECD Declaration and Decisions on International Investment and Multinational Enterprises*. https://www.oecd.org/investment/investment-policy/oecddeclarationanddecisions.htm#:~:text=First%20adopted%20in%201976%2C%20the,to%20economic%20and%20social%20progress, on 8 Mar. 2022.

2011, and 2023.[83] The 2011 version is the first one to explicitly mention compliance programs, although previous versions already mentioned the need for companies to have internal control mechanisms.[84] It does so in the annex of the OECD Guidelines for Multinational Enterprises.[85] This Guidelines address the need for companies to develop and adopt internal controls, ethics, and compliance programs to prevent and detect bribery.

In 2023, an updated version of the 2011 Guidelines, named OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, was published.[86] It encompassed a broader range of forms of corruption, recommending to companies develop and adopt internal controls, ethics, and compliance programs to prevent, detect, and address not only bribery but other forms of corruption.

d. OECD Principles of Corporate Governance (1999, 2004, 2015, 2023)

The 1999 OECD Principles of Corporate Governance was the first initiative by an inter-governmental organization to develop the core elements of a good corporate governance regime.[87] It aimed to be used as a benchmark by governments as they evaluate and improve their laws and regulations, as well as to be helpful to the private sector in developing corporate governance systems and best practices.[88] It stipulates that the corporate governance structure should ensure the integrity of companies' financial and accounting reporting systems and that appropriate control systems should exist, particularly risk monitoring systems, financial controls, and compliance with the law. This version does not even mention the word "corruption."

The 2004 publication provides more detailed guidance on corporate governance structures, advising, for example, that companies, including their subsidiaries, should adopt internal programs and procedures to promote compliance with applicable laws, regulations, and standards, including anti-bribery measures.[89] Therefore, the 2004 OECD Principles of Corporate Governance represent the first version to address anti-corruption compliance programs, although it affirms that factors like the environment, corruption, and ethics are not the central focus of the initiative.

The 2015 version of the OECD Principles of Corporate Governance was designed as an appendix of the Recommendation of the Council on Principles of

---

[83] OECD. *Declaration on International Investment and Multinational Enterprises*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0144, on 27 Dec. 2023.
[84] OECD. (2012). *The OECD Declaration and Decisions on International Investment and Multinational Enterprises: Basic Texts.* https://www.oecd.org/daf/inv/investment-policy/ConsolidatedDeclarationTexts.pdf.
[85] OECD. (2011). *Guidelines for Multinational Enterprises.* https://www.oecd.org/daf/inv/mne/48004323.pdf.
[86] OECD. (2023). *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*. https://www.oecd-ilibrary.org/deliver/81f92357-en.pdf?itemId=/content/publication/81f92357-en&mimeType=pdf.
[87] OECD. (1999). *OECD Principles of Corporate Governance.* https://one.oecd.org/document/C/MIN(99)6/En/pdf.
[88] OECD. (1999). *OECD Principles of Corporate Governance.* https://one.oecd.org/document/C/MIN(99)6/En/pdf.
[89] OECD. (2004). *OECD Principles of Corporate Governance*. https://www.oecd.org/corporate/ca/corporategovernanceprinciples/31557724.pdf.

Corporate Governance.[90] However, this time, the principles were endorsed by the G20 and were referred to as the G20/OECD Principles of Corporate Governance.[91] Regarding compliance programs, it is similar to the 2004 version, although it recommends that the scope of compliance should expand to cover other regulations like taxation, human rights, environment, fraud, and money laundering. Moreover, the 2015 version not only mentions extending compliance programs to subsidiaries but also recommends extending them to third parties, highlighting the importance of monitoring the actions of external entities representing the company.

In 2023, an updated version of the Recommendation of the Council on Principles of Corporate Governance was published.[92] The main objective of this revision was to encourage corporate governance policies that foster the sustainability and resilience of corporations, thereby potentially benefiting the overall economy. There were no changes regarding compliance programs compared to the 2015 version.

e. OECD Recommendation of the Council on Public Procurement (2008, 2015)

This Recommendation has two versions, one from 2008[93] and another from 2015.[94] Its objective is to assist states in promoting appropriate measures for preventing corruption in public procurement. According to the OECD, improving the public procurement process is essential as it represents a massive portion of expenditures, is a crucial pillar of governance and public service delivery, and serves as a tool for achieving pressing political objectives. In order to support states in implementing the guidance of the Recommendation, the OECD has published related documents, such as the Checklist for supporting the implementation of the OECD Recommendation of the Council on Public Procurement[95] and the Public Procurement Toolbox.[96]

The 2008 Recommendation encourages countries to foster close cooperation between the government and the private sector in order to maintain high standards of integrity,[97] particularly in the management of contracts related to public procurement. However, there is no specific provision regarding compliance programs. The 2015 Recommendation explicitly guides that states preserve the integrity of the public

---

[90] OECD. (2015) *Recommendation of the Council on Principles of Corporate Governance*. https://legalinstruments.oecd.org/api/download/?uri=/private/temp/8e37cf70-2ca3-46a6-a340-a2acdf571752.pdf&name=OECD-LEGAL-0413-en%20(2015%20version).pdf.

[91] OECD. (2015). *G20/OECD Principles of Corporate Governance*. https://www.oecd.org/corporate/ca/Corporate-Governance-Principles-ENG.pdf.

[92] OECD. (2023). *Recommendation of the Council on Principles of Corporate Governance.* https://legalinstruments.oecd.org/public/doc/322/322.en.pdf.

[93] OECD. (2008). *Recommendation on Enhancing Integrity in Public Procurement*. https://one.oecd.org/document/C(2008)105/en/pdf.

[94] OECD. (2015). *Recommendation of the Council on Public Procurement.* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0411.

[95] OECD. (2016) *Checklist for Supporting the Implementation of the OECD Recommendation of the Council on Public Procurement.* https://www.oecd.org/gov/ethics/High-Level_Principles_Integrity_Transparency_Control_Events_Infrastructures.pdf.

[96] OECD. *Public Procurement Toolbox.* https://www.oecd.org/governance/procurement/toolbox/.

[97] The OECD defines integrity as "The consistent alignment of, and adherence to, shared ethical values, principles, and norms for upholding and prioritizing the public interest over private interests." (OECD. (2019). *Recommendation of the Council on Guidelines on Anti-corruption and Integrity in State-owned Enterprises*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0451, at 5.)

procurement system by requiring the private sector to adopt internal controls, compliance standards, and anti-corruption programs. The state should monitor the implementation of such mechanisms by companies. The Recommendation also stipulates that contracts for public procurement should include guarantees of "non-corruption" by the private sector, which the state should verify, and contractors should be encouraged to promote transparency and provide integrity training to members of their supply chains, aiming to combat corruption in subcontracting as well.

3. United Nations

The United Nations (UN) is an international organization founded in 1945 and currently composed of 193 member states.[98] Targeting member states, the UN's most famous anti-corruption instrument is the international treaty United Nations Convention Against Corruption (hereafter UN Convention). The UN also provides guidelines aiming to implement anti-corruption strategies in their own operations.[99]

a. UN Convention (2003)

The UN Convention was adopted by the UN General Assembly in 2003 and entered into force in 2005.[100] The UN Convention is considered the only legally binding and universal anti-corruption instrument.[101] Currently, the UN Convention has 189 state parties, making it the most subscribed international anti-corruption instrument.[102] The UN Convention is also the most comprehensive legal instrument in the international anti-corruption law realm, covering five main areas: (i) preventive measures, (ii) criminalization and law enforcement, (ii) international cooperation, (ii) asset recovery, and (iii) technical assistance and information exchange.[103] Like the OAS and OECD Conventions, it also urges states to hold companies liable for the illicit acts outlined within it.[104] The treaty emphasizes the importance of private sector

---

[98] UN. *About us*. https://www.un.org/en/about-us, on 1 Mar. 2022. Its objectives are listed in its founding Charter and include promoting social progress and better standards of life in larger freedom. The Charter does not specify the fight against corruption as a UN objective (UN. *United Nations Charter (full text)*. https://www.un.org/en/about-us/un-charter/full-text, on 1 Mar. 2022.)

[99] Chapters II, III, IV, V, and VI, respectively. UN. (2016). *Comprehensive Review of Governance and oversight within the United Nations and its Funds, Programmes and Specialized Agencies.* https://undocs.org/en/A/RES/61/245; UN. (2020). *Enterprise Risk Management: Approaches and Uses in United Nations System Organizations*. https://www.unjiu.org/sites/www.unjiu.org/files/jiu_rep_2020_5_english.pdf.

[100] UN. *United Nations Convention Against Corruption.* https://www.unodc.org/unodc/en/corruption/uncac.html, on 1 Mar. 2022.

[101] UN. *United Nations Convention Against Corruption.* https://www.unodc.org/unodc/en/corruption/uncac.html, on 4 Jul. 2023.

[102] UN. *Signature and Ratification Status*. https://www.unodc.org/unodc/en/corruption/ratification-status.html, on 4 Jul. 2023.

[103] UN. (2004). *United Nations Convention Against Corruption*. https://www.unodc.org/documents/treaties/UNCAC/Publications/Convention/08-50026_E.pdf; STAR. *About UNCAC.* https://star.worldbank.org/focus-area/uncac#:~:text=The%20Convention%20covers%20five%20main,technical%20assistance%20and%20information%20exchange, on 1 Mar. 2022.

[104] Article 26 (UN. (2004). *United Nations Convention Against Corruption*. https://www.unodc.org/documents/treaties/UNCAC/Publications/Convention/08-50026_E.pdf). The document On the Level: Business and Governments Against Corruption summarizes the UN Convention recommendations regarding the fight against corruption in the private sector (UNODC. *On the Level: Business and Governments Against Corruption*.

involvement in fighting corruption and urges countries to promote this engagement, for instance, by enhancing account and audit standards in the private sector,[105] however it does not explicitly refer to compliance programs. Other documents produced by the UN after the Convention do, as shown below.[106]

While the binding nature of the UN Convention allows states parties significant discretion in determining how it affects their legal systems, as "the provisions fall into many gradations, ranging from mandatory to non-mandatory, precise to vague, and absolute to qualified."[107] Nevertheless, these provisions can still drive domestic anti-corruption measures, influencing discussions on what actions should be criminalized and how to address them within domestic policy.[108] The UN Conventions did not set forth a follow-up mechanism. However, in 2009, the UN established the Mechanism for the Review of Implementation of the UN Convention to assist members in implementing the treaty.[109]

b. Business Against Corruption: A Framework for Action (2005, 2011)

This document results from a collaboration between the UN Global Compact,

---

https://www.unodc.org/documents/corruption/Publications/2014/UNODC_On_the_Level_Business_and_Government_against_Corruption.pdf, on 8 Aug. 2023).

[105] *See* Article 12 (UN. (2004). *United Nations Convention Against Corruption*. https://www.unodc.org/documents/treaties/UNCAC/Publications/Convention/08-50026_E.pdf). The document On the Level: Business and Governments Against Corruption summarizes the UN Convention recommendations regarding the fight against corruption in the private sector (UNODC. *On the Level: Business and Governments Against Corruption*. https://www.unodc.org/documents/corruption/Publications/2014/UNODC_On_the_Level_Business_and_Government_against_Corruption.pdf, on 8 Aug. 2023).

[106] The UN website's search mechanism lacks filter options, leading to challenges in searching for instruments related to compliance programs. The results are extensive and encompass diverse types of documents, including those irrelevant to this article. For instance, a search on August 8, 2023, for "compliance program" yielded 1,184 results (UN. *Site search*. https://www.un.org/site-search/). For this reason, I searched the UN documents presented here through two other approaches. Firstly, I examined UN documents referenced within the materials of other investigated international actors. Secondly, I navigated the "Documents, publications and tools" page of the UNODC site, scrutinizing those listed under the "Corruption and the private sector" section. The following resources were on this list: (1) UNODC Business Integrity Portal; (2) UNODC-UN Global Compact anti-corruption e-learning tool for the private sector; (3) Anti-Corruption Policies and Measures of the Fortune Global 500; (4) Corruption Prevention to Foster SME Development (UNIDO/UNODC - 2 volumes) Volume 1 (English) - Volume 2 (English) - Volume 1 (Spanish); (5) An Anti-Corruption Ethics and Compliance Programme for Business: A Practical Guide English - French - Spanish – Russian; (6) Anti-Corruption Ethics and Compliance Handbook for Business (OECD/UNODC/World Bank) English – Spanish; (7) Corporate Integrity: Incentives for Corporate Integrity in Accordance with the United Nations Convention Against Corruption – A Report; (8) A Resource Guide on State Measures for Strengthening Corporate Integrity English - Spanish – Russian; (9) On the Level: Business and Governments Against Corruption, Toolkit of Private Sector Outreach Materials; (10) The Puppet Masters: How the Corrupt Use Legal Structures to Hide Stolen Assets and What to Do About It. (UNODC. *Documents, Publications and Tools – Corruption and the Private Sector*. https://www.unodc.org/unodc/en/corruption/publications.html, on 8 Aug. 2023). In this section, I described these documents, except for the sixth one, which is detailed in another section of this work titled "World Bank Framework," and the tenth, the link for which on the UNODC site leads to a "Page Not Found." (https://star.worldbank.org/star/publication/puppet-masters, on 8 Aug. 2023).

[107] CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015), at 97.

[108] CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015).

[109] UNODC. (2011). *Mechanism for the Review of Implementation of the United Nations Convention Against Corruption – Basic Documents*.

TI, and the International Business Leaders Forum. It has two editions, one from 2005[110] and another from 2011.[111] Both editions of the guidance highlight the importance of the private sector in addressing corruption, emphasizing why corruption is detrimental to business and encouraging compliance programs. The 2005 edition states that the first step for companies to comply with the tenth principle of the UN Global Compact is to introduce anti-corruption policies and programs within their organizations and business operations. Among the recommendations, the document suggests that companies, regardless of size, adopt the TI Six-Step Implementation Process, a process for developing and implementing an anti-bribery policy.[112]

The 2011 document acknowledges that while there has been progress in the global fight against corruption since 2005, the problem persists. To address corruption in companies, in addition to the compliance programs, the guide suggests adopting the UN Global Compact Management,[113] a management tool produced by the UN Global Compact and Deloitte aimed at corporate sustainability.[114] Moreover, the 2011 version mentions the importance of collective action, defined in the document as a cooperative process among various stakeholders to combat corruption jointly, allowing for the alliance of organizations with similar objectives but with different perspectives on the problem, enabling new solutions that increase the impact of individual actions. The guide also states that the ultimate goal of these joint efforts should be to create fair and equal conditions for all market participants and to eliminate corruption temptations for everyone.

c. Corruption Prevention to Foster Small and Medium-Sized Enterprise Development (2007, 2012)

The United Nations Industrial Development Organization (UNIDO) and the

---

https://www.unodc.org/documents/treaties/UNCAC/Publications/ReviewMechanism-BasicDocuments/Mechanism_for_the_Review_of_Implementation_-_Basic_Documents_-_E.pdf

[110] UN Global Compact, TI, International Business Leaders Forum. (2005). *Business Against Corruption: A Framework for Action – Implementation of the 10th UN Global Compact Principle Against Corruption*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2F7.7%2FBACtextcoversmallFINAL.pdf.

[111] UN Global Compact, TI, International Business Leaders Forum. (2011). *Business Against Corruption: A Framework for Action*. https://d306pr3pise04h.cloudfront.net/docs/news_events%2F8.1%2Fbac_fin.pdf.

[112] "Transparency International has developed a Six-Step Implementation Process based on the Business Principles for Countering Bribery. This practical guide assists companies in developing and implementing an anti-bribery policy. The TI Six-Step Implementation Process can be modified to take into account the size of a company and its ability to complete the steps within the suggested timeframe." (UN Global Compact, TI, International Business Leaders Forum. (2005). *Business Against Corruption: A Framework for Action – Implementation of the 10th UN Global Compact Principle Against Corruption*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2F7.7%2FBACtextcoversmallFINAL.pdf, at 11).

[113] UN Global Compact, Deloitte. (2010). *UN Global Compact Management Model*. https://d306pr3pise04h.cloudfront.net/docs/news_events%2F9.1_news_archives%2F2010_06_17%2FUN_Global_Compact_Management_Model.pdf.

[114] "A practical yet comprehensive tool to help companies evolve their sustainability efforts. Comprised of six management steps, it guides companies of all sizes through the process of formally committing to, assessing, defining, implementing, measuring and communicating a corporate sustainability strategy. The model draws on widely accepted and understood management practices, and is designed to maximize corporate sustainability performance." (UN. *UN Global Compact Management Model.* https://www.unglobalcompact.org/library/231, on 14 Dec. 2021).

United Nations Office on Drugs and Crime (UNODC)[115] have a project on Corruption Prevention to Foster Small and Medium Sized Enterprise Development, aimed to develop a service concept and related tools to support small and medium-sized enterprises (SMEs) in defending themselves against both public and private sector corruption.[116] The project produced two reports.

The first report, published in 2007, primarily analyzes the challenges posed to SME development by public and private sector corruption.[117] It also discusses potential measures and tools to assist these companies in combatting corruption in their operations. The document emphasizes that although SMEs frequently implement internal measures more readily and swiftly than larger companies, relying solely on compliance programs may not be effective for them. This is often due to their limited resources and market influence, which can hinder their ability to uphold zero-tolerance corruption policies. Furthermore, SMEs face the risk of losing market share to competitors that do not adhere to such standards. Therefore, the report recommends implementing additional measures to empower SMEs against corruption, such as collective actions. The 2012 version restates the first one regarding compliance programs, focusing and deepening on tools and measures other than compliance programs to support SMEs in their fight against corruption.[118]

d. Global Compact for the 10th Principle (2009, 2012)

The UN Global Compact, launched in 2000, is a call for companies to align their strategies and operations with ten principles.[119] This initiative is recognized as the world's largest corporate sustainability initiative, with over 16,000 members in 160 countries.[120] Such guidance aims to promote the 10th principle of the UN Global Compact, the fight against corruption by the private sector, by providing a roadmap of

---

[115] UNODC is the leading entity in the fight against corruption, the guardian of the UN Convention, and the guardian of the Global Compact 10th Principle (UNODC. *UNODC Business Integrity Portal.* https://businessintegrity.unodc.org/, on 7 Aug. 2023).

[116] UNIDO. *Corruption Prevention to Foster Small and Medium Sized Enterprise Development*. https://www.unido.org/our-focus/advancing-economic-competitiveness/competitive-trade-capacities-and-corporate-responsibility/corporate-social-responsibility-market-integration/csr-projects/corruption-prevention-foster-sme-development-unido-joins-forces-unodc, on 8 Aug. 2023.

[117] UNIDO, and UNODC. (2007). *Corruption Prevention to Foster Small and Medium-Sized Enterprise Development: Providing Anti-Corruption Assistance to Small Businesses in The Developing World.* https://www.unodc.org/documents/corruption/Publications/2012/UNIDO-UNODC_Publication_on_Small_Business_Development_and_Corruption_Vol1.pdf

[118] UNIDO, and UNODC. (2012). *Corruption Prevention to Foster Small and Medium-Sized Enterprise Development.* https://www.unodc.org/documents/corruption/Publications/2012/Corruption_prevention_to_foster_small_and_medium_size_enterprise_development_Vol_2.pdf

[119] The Global Compact is an initiative by the UN that aims to "Accelerate and scale the global collective impact of business by upholding the Ten Principles and delivering the SDGs through accountable companies and ecosystems that enable change. To make this happen, the UN Global Compact supports companies to: 1. Do business responsibly by aligning their strategies and operations with Ten Principles on human rights, labour, environment and anti-corruption; and 2. Take strategic actions to advance broader societal goals, such as the UN Sustainable Development Goals, with an emphasis on collaboration and innovation." (UN Global Compact. *Who we are*. https://unglobalcompact.org/what-is-gc/mission, on 8 Aug. 2023).

[120] UN Global Compact. *Who we are*. https://unglobalcompact.org/what-is-gc/mission, on 8 Aug. 2023

resources and tools to assist companies in anti-corruption actions.[121] The 10th Principle of the UN Global Compact focuses on combating corruption, and about this principle,[122] the Global Compact has issued two documents. One, published in 2009, named Reporting Guidance on the 10th Principle Against Corruption published, in collaboration with TI.[123] Another, published in 2012, named Global Compact for the 10th Principle: Corporate Sustainability with Integrity – Organizational Change to Collective Action.[124]

The 2009 report clarifies that the adoption of the 10th Principle commits participants not only to avoid bribery, extortion, and other forms of corruption, but also to develop policies and concrete programs to address it. This report also provides a set of elements to help any organization identify the components of a comprehensive anti-corruption program. The 2012 document expands on the recommendations related to compliance programs. It affirms that the Global Compact addresses corruption by advocating for stringent anti-corruption practices through both individual company-level changes and collaborative efforts at the national level. Companies are urged to integrate anti-corruption measures into their strategies and operations, involving codes of conduct, zero-tolerance policies, and regulations on various aspects such as gifts, politics, and travel. The report recommends that actions like anonymous hotlines, training, supply chain management, risk assessment, and disciplinary measures should support anti-corruption measures. Moreover, the report encourages collective actions.

e. Fighting Corruption in the Supply Chain: A Guide for Customers and Suppliers (2010, 2016)

The Global Compact also produced the guide Fighting Corruption in the Supply Chain, which has two versions: one from 2010[125] and another from 2016.[126] Both versions aim to guide the private sector in combating corruption in their supply chains, recognizing that most companies are both customers as well as suppliers. The documents state that tackling corruption in the supply chain should be part of a broader anti-corruption program that addresses corruption risks throughout the company, regardless of its size and scope. Both versions of the guide affirm that for all companies, combating corruption in the supply chain must be part of a larger anti-corruption

---

[121] "Principle 10: Businesses should work against corruption in all its forms, including extortion and bribery." (UN. *The Ten Principles of the UN Global Compact*. https://unglobalcompact.org/what-is-gc/mission/principles, on 7 Aug. 2023).

[122] UN. *Global Compact.* https://unglobalcompact.org/, on 7 Aug. 2023.

[123] UN and TI. (2009) *Reporting Guidance on the 10th Principle Against Corruption*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FUNGC_AntiCorruptionReporting.pdf.

[124] UN Global Compact. (2012). *Global Compact for the 10th Principle: Corporate Sustainability with Integrity –Organizational Change to Collective Action.* https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FGC_for_the_10th_Principle.pdf.

[125] UN. (2010). *Fighting Corruption in The Supply Chain: A Guide for Customers and Suppliers.* https://www.unglobalcompact.bg/wp-content/uploads/2014/05/131.pdf.

[126] UN. (2016). *Fighting Corruption in The Supply Chain: A Guide for Customers and Suppliers.* https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FFighting_Corruption_Supply_Chain.pdf.

program that addresses corruption risks throughout the firm.

f. An Anti-Corruption Ethics and Compliance Programme for Business: A Practical Guide (2013)

In 2013, the UNODC produced a guide to assist businesses in implementing an effective anti-corruption ethics and compliance program.[127] The document highlights several reasons to demonstrate that corruption is bad for business and emphasizes that the adoption of a compliance program adds value to the company. It also states that the program not only promotes adherence to laws but also serves as a crucial element in protecting the company's reputation and the interests of investors and shareholders. Furthermore, it affirms that the adoption of a compliance program is good for all businesses as it contributes to a fair market without distortions caused by corrupt practices.

g. A Resource Guide on State Measures for Strengthening Corporate Integrity (2013)

This document, produced by UNODC, primarily aims to explore measures that countries can use to promote corporate integrity.[128] The document is structured into three sections. The first segment begins by detailing the articles of the UN Convention that provide a framework for states' engagement with the private sector. The second section presents a business case for combating corruption and the core elements of an effective anti-corruption program. The final section describes the range of sanctions and incentives that states can be used to advance the UN Convention's objectives for preventing and addressing corruption within the private sector.

The guide affirms that the implementation of a meaningful and effective anti-corruption program for business is primarily a private sector function and responsibility. However, "corporate anti-corruption programs are a primary tool for strengthening integrity and should be encouraged."[129] Thus, states should help shape the corporate investment decision for the implementation of a compliance program through a combination of enforcement sanctions and good practice incentives. As incentives that the states can give, the guide listed: penalty mitigation to encourage self-reporting of offenses and give credits to company-led prevention efforts; procurement incentives to reward good practice through procurement preference; preferential access to government benefits to reward good practice with preferential access to government, making, for instance, the access to government support or services conditional on minimum integrity practices; reputational benefits to encourage good practice through public recognition; and whistleblower awards to promote reporting of potential violations by individuals.

The guide also urges states to endorse additional measures to promote integrity,

---

[127] UNODC. (2013). *An Anti-Corruption Ethics and Compliance Programme for Business: A Practical Guide* https://www.unodc.org/documents/corruption/Publications/2013/13-84498_Ebook.pdf.

[128] UNODC. (2013). *A Resource Guide on State Measures for Strengthening Corporate Integrity*. https://www.unodc.org/documents/corruption/Publications/2013/Resource_Guide_on_State_Measures_for_Strengthening_Corporate_Integrity.pdf.

[129] UNODC. (2013). *A Resource Guide on State Measures for Strengthening Corporate Integrity*. https://www.unodc.org/documents/corruption/Publications/2013/Resource_Guide_on_State_Measures_for_Strengthening_Corporate_Integrity.pdf, at 2.

aligning with the private sector, such as collective acts, such as integrity pacts. These pacts serve as mechanisms for elevating integrity standards within projects or sectors, achieved through contractual commitments and third-party supervision. Furthermore, states can promote initiatives based on codes of conduct involving businesses to enhance awareness and fortify integrity practices on local, regional, or sectoral levels. Moreover, states can undertake public sector reforms to stimulate collaborative public-private endeavors targeting the corruption demand side, facilitated by civil services and regulatory adjustments.

h. Connecting the Business and Human Rights and the Anti-corruption Agendas (2020)[130]

This report by the Working Group on Human Rights and Transnational Corporations explores the intersection of corruption and human rights within business-related activities.[131] The report highlights the potential synergy between measures promoting responsible business practices, human rights, and anti-corruption efforts to reinforce each other to ensure a corporate coherent policy. The document underscores that while companies have implemented anti-corruption compliance programs to manage risks, such efforts often neglect human rights considerations due to the absence of regulatory requirements. It points out challenges to the integration, such as different department for sustainability, human rights, and anti-corruption in the companies, hindering effective communication and collaboration.

The document highlights that key actor, such as the PACI and Global Compact, have called for a holistic, integrated approach to responsible business conduct. However, despite the expectations set by these actors, businesses lag behind in implementing human rights due diligence processes alongside existing integrity and anti-corruption measures. Its affirms that not many companies are genuinely focusing on such alignment. The document highlights good practices to do that, such as emphasizing a culture of integrity led by senior business leaders, considering corruption and human rights risks in employee onboarding, adopting standard codes of conduct with clauses on human rights and anti-corruption, covering human rights and corruption in non-financial audits, integrating human rights into anti-corruption training, aligning the identification of human rights and anti-corruption risks, and incorporating corruption risks into human rights due diligence through the compliance department.

The Working Group on Human Rights and Transnational Corporations also urges states to translate anti-corruption policies into action, addressing business-related human rights impacts through responsible business conduct. Recommendations for states include providing technical assistance to those lacking capacity, breaking institutional silos, introducing regulations mandating human rights due diligence, examining integrity and anti-corruption pledges with an expanded focus on human rights, reviewing withdrawal of support from companies engaged in bribery or corruption, promoting policy coherence in combating corruption and human rights

[130] I did not find this document through a search on the UN page for it name; instead, it was identified because the preliminary notes that led to this document were referenced in one of the PACI documents analyzed in this article, *see* WEF. (2020). *Agenda for Business Integrity: Collective Action – Community Paper*. https://www3.weforum.org/docs/WEF_Agenda_for_Business_Integrity.pdf.

[131] UN. (2020). *Connecting the Business and Human Rights and the Anti-Corruption Agendas: Report of the Working Group on the Issue of Human Rights and Transnational Corporations and other Business Enterprises (A/HRC/44/43)*. https://digitallibrary.un.org/record/3889182, on 6 Dec. 2023.

abuses, and enhancing integrity pact processes to monitor business respect for human rights. It also calls on civil society to promote this holistic corporate culture by, for instance, documenting cases, engaging in collective action, and advocating for innovative anti-corruption mechanisms.

i. Other UN initiatives

In 2013, the UNODC produced a document analyzing India's progress in promoting corporate integrity in alignment with the principles of the UN Convention. The document, titled Corporate Integrity: Incentives for Corporate Integrity in Accordance with the United Nations Convention Against Corruption – A Report, acknowledges that fines are generally the most frequently used sanction against corruption, followed by exclusion from government contracts, forfeiture, confiscation, restitution, debarment, or legal entity closure.[132] Additionally, it recommends that states consider non-monetary sanctions for corruption, including the establishment of effective internal compliance programs and direct regulation of corporate structures.

The UNODC also produced resources to improve corporate integrity by targeting companies. The UNODC Business Integrity Portal is a "platform for strengthened dialogue and partnership between the public and the private sectors to develop and implement initiatives to counter corruption jointly."[133] The site offers several resources to aid companies in developing compliance programs and other measures against corruption. For instance, the Business Hub "offers information and tools to businesses seeking to strengthen integrity in their operations by assessing their corruption risks, developing compliance programs, and participating in Collective Action activities."[134] The UNODC also produced the Toolkit of Private Sector Outreach Materials, a summary of documents produced by the UN.[135] The UN, by UNODC e Global Compact, also provides The Fight Against Corruption, an e-learning tool for the private sector in the UN Global Compact's 10th Principle.[136]

B. Intergovernmental Initiatives

1. G20

The G20 was established in 1999 and recognized in 2009 as the foremost forum for international economic collaboration.[137] It currently comprises 19 countries and the

---

[132] UNODC. (2013). *Corporate Integrity: Incentives for Corporate Integrity in Accordance with the United Nations Convention Against Corruption – A Report*. https://www.unodc.org/documents/southasia//publications/research-studies/CI_Report.pdf.

[133] UNODC. *UNODC Business Integrity Portal.* https://businessintegrity.unodc.org/bip/en/index.html, on 8 Aug. 2023.

[134] UNODC. *Business Hub*. https://businessintegrity.unodc.org/bip/en/business-hub.html, on 8 Aug. 2023.

[135] UNODC. *Toolkit of Private Sector Outreach Materials*. https://www.unodc.org/unodc/en/corruption/tools_and_publications/toolkit-of-private-sector-outreach-materials.html, on 8 Aug. 2023.

[136] UNODC, and Global Compact. *The Fight Against Corruption.* http://thefightagainstcorruption.org/, on 8 Aug. 2023.

[137] Established in 1999 following the Asian financial crisis, the G20 emerged as a platform for Finance Ministers and Central Bank Governors to deliberate on global economic and financial matters. In response to the worldwide economic and financial crisis of 2007, the group included the Heads of Government. By 2009, it gained recognition as the foremost forum for international economic

European Union.[138] Initially centered on broader macroeconomic concerns, the G20 has progressively broadened its scope to encompass various subjects, including anti-corruption efforts.[139] None of Villarino, Rose, or WRC included the G20 initiatives within their listings of the IACR.[140] However, I chose to include the G20 in this article because both the OECD and UN have referenced G20 anti-corruption standards in some documents analyzed in this article. Furthermore, G20 actions can significantly stimulate anti-corruption efforts, given that its members represent around 85% of the global Gross Domestic Product (GDP), over 75% of global trade, and about two-thirds of the world's population.[141] Consequently, I have undertaken an analysis of the G20 instruments below.[142]

a. G20 ACWG Action Plan (2011-2012, 2013-2014, 2015-2016, 2017-2018, 2019-2021, 2022-2024)

The G20 Anti-Corruption Working Group (G20 ACWG), established in 2010, has the specific objective of preparing comprehensive recommendations for consideration by the leaders of G20 member countries on how to contribute to international efforts to combat corruption.[143] Since 2011, the G20 ACWG has been

collaboration. (G20. *About G20 – Overview.* https://www.g20.in/en/about-g20/about-g20.html#overview, on 10 Aug. 2023).

[138] Argentina, Australia, Brazil, Canada, China, France, Germany, India, Indonesia, Italy, Japan, Republic of Korea, Mexico, Russia, Saudi Arabia, South Africa, Türkiye, United Kingdom, and the United States. (G20. *About G20 – G20 Members.* https://www.g20.org/en/about-g20/#members, on 10 Aug. 2023).

[139] G20. *About G20.* https://www.g20.in/en/about-g20/about-g20.html, on 10 Aug. 2023.

[140] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL, 343 (2022); CECILY ROSE, INTERNATIONAL ANTI-CORRUPTION NORMS (2015); Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[141] OECD and UNDP. (2019). *G20 Contribution to the 2030 Agenda Progress and Way Forward*. www.oecd.org/dev/OECD-UNDP-G20-SDG-Contribution-Report.pdf.

[142] On February 20, 2022, in my search for G20-produced documents about anti-corruption, I founded the "G20 Anti-Corruption Resources", a virtual library hosted on the UNODC website (https://www.unodc.org/unodc/en/corruption/g20-anti-corruption-resources/by-thematic-area.html). This repository encompassed all G20 ACWG Action Plans (2011-2012, 2013-2014, 2015-2016, 2017-2018, 2019-2021, 2022-2024). Furthermore, the library provided access to various other G20 anti-corruption instruments. Due to the considerable volume of materials available and the objective of this article, I reviewed the titles of all the documents listed in the virtual library. I selected those related to the prohibition of corrupt activities and those concerning the private sector. I selected the following papers: 1) G20 Guiding Principles on Enforcement of the Foreign Bribery Offence (2013); (2) G20 Guiding Principles to Combat Solicitation (2013); (3) G20 High Level Principles on Corruption and Growth (2014); (4) G20 High-Level Principles on Beneficial Ownership Transparency (2014); (5) G20 High-Level Principles on Private Sector Transparency and Integrity (2015); (6) G20 Principles for Promoting Integrity in Public Procurement; (7) G20 High Level Principles on the Liability of Legal Persons for Corruption (2017); (8) G20 High Level Principles on Organizing Against Corruption (2017); (9) G20 High-Level Principles for Preventing Corruption and Ensuring Integrity in State-Owned Enterprises (2018); (10) G20 Compendium of Good Practices for Promoting Integrity and Transparency in Infrastructure Development (2019); (11) G20 High-Level Principles for the Development and Implementation of National Anti-Corruption Strategies. I described the documents that mention compliance programs in this section.

[143] STAR. *G20 Anti-Corruption Working Group*. https://star.worldbank.org/g20-anti-corruption-working-group, on 20 Feb. 2022.

publishing a multi-year anti-corruption plan.[144] These plans, built upon the conventions to which G20 members have signed, like the OECD and the UN Convention, and the monitoring and accountability reports to the G20, set priority goals for the corresponding period. A common theme in all the plans is the need to engage the private sector in the fight against corruption, with partnerships between governments and businesses viewed as essential in addressing the problem.

The first plan that explicitly addresses compliance programs is the plan for the 2015-2016 biennium, stating that states should encourage the private sector to adopt robust compliance programs.[145] The two subsequent plans, 2017-2018[146] and 2019-2021[147] do not explicitly mention compliance programs. However, they emphasize the G20's commitment to fostering a corporate culture of integrity and endorsing private-sector anti-corruption initiatives. The G20 Anti-corruption Action Plan 2022-2024 speaks once more specifically about compliance programs, highlighting that the G20 will continue to encourage and support efforts by the private sector to strengthen effective internal controls and anti-corruption ethics and compliance programs.[148]

b. G20 Principles for Promoting Integrity in Public Procurement (2015)

Similarly to the OECD in the 2015 Recommendation of the Council on Public Procurement described above,[149] these principles, also published in 2015, recognize

---

[144] UNODC. (2010). *G20 Anti-corruption Action Plan 2011-2012*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2010_G20_ACWG_Action_Plan_2011-2012.pdf; UNODC. (2012). *G20 Anti-Corruption Action Plan 2013-2014*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2012_G20_ACWG_Action_Plan_2013-2014.pdf; UNODC. (2014). *G20 Anti-corruption Action Plan 2015-2016*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2014_G20_ACWG_Action_Plan_2015-2016.pdf; UNODC. (2015). *G20 Anti-corruption Action Plan 2017-2018*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2015_G20_ACWG_Action_Plan_2017-2018.pdf; UNODC. (2016). *G20 Anti-Corruption Action Plan 2019-2021*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2018_G20_ACWG_Action_Plan_2019-2021.pdf; UNODC. (2021). *G20 Anti-corruption Action Plan 2022-2024*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2021_G20_Anti-Corruption_Action_Plan_2022-2024.pdf.

[145] UNODC. (2014). *G20 Anti-corruption Action Plan 2015-2016*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2014_G20_ACWG_Action_Plan_2015-2016.pdf.

[146] *G20 Anti-corruption Action Plan 2017-2018*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2015_G20_ACWG_Action_Plan_2017-2018.pdf.

[147] UNODC. (2016). *G20 Anti-Corruption Action Plan 2019-2021*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2018_G20_ACWG_Action_Plan_2019-2021.pdf.

[148] UNODC. (2021). *G20 Anti-corruption Action Plan 2022-2024*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2021_G20_Anti-Corruption_Action_Plan_2022-2024.pdf.

[149] OECD. (2015). *Recommendation of the Council on Public Procurement.* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0411.

the high risks of corruption in public procurement.[150] This G20 document states that, given the vast resources and close interaction between the public and private sectors, public procurement processes are particularly vulnerable to corruption and other misconduct, leading to inefficient allocation of public resources and decreased citizens' trust in good governance. As one of the strategies to address the problem, the document suggests that states promote a culture of integrity by encouraging contractors to develop internal controls and compliance standards, including anti-corruption programs, and find ways to give proper recognition to contractors with effective anti-corruption mechanisms in place.

c. G20 High-Level Principles on Private Sector Transparency and Integrity (2015)

Unlike the other G20 documents presented here, which are addressed to states, the G20 High-Level Principles on Private Sector Transparency and Integrity, published in 2015, are directed toward companies.[151] The document asserts that the G20 seeks to encourage the commitment of companies, ranging from small enterprises to large corporations, to improve internal controls, ethics and compliance, transparency, and integrity. The document also affirms that the G20 will continue to collaborate with companies and other stakeholders, including the B20[152] and C20,[153] to promote compliance through collective action and public-private dialogues, as well as support the development and implementation of anti-corruption programs in companies.

d. G20 High-Level Principles on the Liability of Legal Persons for Corruption (2017)

Dated as of 2017, this document established principles aimed at identifying mechanisms and practices useful for states in establishing and enforcing the liability of legal persons for corruption and related offenses.[154] The 13th principle guides states to promote the private sector to develop anti-corruption actions. The 14th principle asserts

---

[150] UNODC. (2015). *G20 Principles for Promoting Integrity in Public Procurement*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Public-Sector-Integrity-and-Transparency/G20-Principles_for_Promoting_Integrity_in_Public_Procurement_2015.pdf.

[151] UNODC. (2015). *G20 High-Level Principles on Private Sector Transparency and Integrity.* https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Private-Sector-Integrity-and-Transparency/G20_High_Level_Principles_on_Private_Sector_Transparency_and_Integrity_2015.pdf.

[152] "The Business 20 (B20) represents the voice of business in the G20, an intergovernmental forum representing the world's major economies. It acts as a platform for dialogue between businesses and the G20 presidency, which rotates each year, with input from civil society and international organisations. [...] Given the B20's influence on both government policies and the private sector, it is crucial that issues of business ethics and integrity are central to B20 recommendations." (Basel Institute on Governance. *B20 and Anti-corruption*. https://baselgovernance.org/b20-collective-action-hub/b20-anti-corruption, on 1 Mar. 2022).

[153] "C20 is one of the official Engagement Groups of the G20. It provides a platform for Civil Society Organizations (CSO) around the world to bring forth the political dialogue with the G20. The Civil 20 process involves a wide variety of organizations and networks far beyond the G20 countries and it is structured around the C20 Guiding Principles." (C20. *About C20*. https://civil-20.org/index.php/about-c20/, on 1 Mar. 2022).

[154] UNODC. (2017). *G20 High-Level Principles on the Liability of Legal Persons for Corruption*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Bribery/G20_High_Level_Principles_on_the_Liability_of_Legal_Persons_for_Corruption_2017.pdf.

that the state may, for this purpose, consider the existence of corporate anti-corruption ethics and compliance programs or measures in decisions regarding public procurement or other processes aimed at granting public benefits, such as export credits. Moreover, this instrument states that governments should recognize the efforts made by companies to develop and implement such anti-corruption strategies. In addition, it suggests that when appropriate and consistent with the member country's legal system, these efforts can be considered in judicial proceedings, for example, as a mitigating factor of the sanction or as a defense.

e. G20 Compendium of Good Practices for Promoting Integrity and Transparency in Infrastructure Development (2019)

The compendium lists best practices in infrastructure development, including compliance programs and guidelines for the private sector.[155] Moreover, it cites Indonesia's anti-corruption action as an example of good practice. Indonesia has developed a voluntary program called the National Movement for Integrity Development in the Business Sector with the slogan "Professional with Integrity – PROFIT." The document also highlights integrity pacts as a good practice. These are a form of collective action aimed at assessing and mitigating corruption risks by the government, businesses, and civil society together.

C. International Financial Institutions

1. World Bank Group

The World Bank Group is an international financial organization that aims to provide sustainable solutions to reduce poverty and build shared prosperity in developing countries, with 189 member countries.[156] The World Bank Group is formed by five institutions: the International Bank for Reconstruction and Development (IBRD); the International Development Association (IDA); the International Finance Corporation (IFC); the Multilateral Investment Guarantee Agency (MIGA); and the International Centre for Settlement of Investment Disputes (ICSID).[157] While the IBRD and IDA – which together constitute the World Bank – provide financing, policy advice, and technical assistance to governments, the IFC, MIGA, and ICSID focus on bolstering the private sector.[158]

The World Bank Group is recognized as a relevant international actor involved

---

[155] UNODC. (2019). *G20 Compendium of Good Practices for Promoting Integrity and Transparency in Infrastructure Development.* https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Sectors/G20_Compendium_of_Good_Practices_for_Promoting_Integrity_and_Transparency_in_Infrastructure_Development_2019.pdf.

[156] World Bank. *Who we are*. https://www.worldbank.org/en/who-we-are, on 11 Aug. 2023. The origins of the World Bank Group can be traced back to 1944, when countries came together to assist in the reconstruction of Europe and Japan following World War II. (World Bank. *Getting to Know the World Bank*, https://www.worldbank.org/en/news/feature/2012/07/26/getting_to_know_theworldbank, on 11 Aug. 2023).

[157] World Bank. *Who we are*. https://www.worldbank.org/en/who-we-are, on 11 Aug. 2023.

[158] World Bank. *Who we are*. https://www.worldbank.org/en/who-we-are, on 11 Aug. 2023.

in anti-corruption activity.[159] In the 1996 Annual Meeting of the World Bank and the IMF, the President of the World Bank Group referred to "the cancer of corruption," a comparison that resonated worldwide and is emblematic of a stand against corruption.[160] Since then, the World Bank Group has taken several actions against corruption. For example, in 2004, they implemented a rule stating that companies bidding on significant projects financed by the Bank must confirm that they have taken measures to prevent any person acting on their behalf from engaging in bribery.[161] In addition, the World Bank Group possesses the power to impose sanctions on companies and individuals that breach its norms in connection with projects financed by a World Bank Group entity.[162] These sanctions complement the IACR.[163]

a. World Bank Group Integrity Compliance Guidelines (2010)

In 2001, the World Bank Group established the Department of Institutional Integrity, currently named the Integrity Vice Presidency (INT),[164] an independent unit that investigates and imposes sanctions related to allegations of fraud and corruption in projects financed by the bank.[165] The sanctions can be imposed by reprimand, conditional non-debarment, debarment, debarment with conditional release, and/or restitution (financial or otherwise), including the temporary or permanent exclusion of

---

[159] Susan Rose-Ackerman, *The Role of International Actors in Fighting Corruption*, in ANTI-CORRUPTION POLICY: CAN INTERNATIONAL ACTORS PLAY A CONSTRUCTIVE ROLE? 3 (Susan Rose-Ackerman & Paul D. Carrington ed., 2014).

[160] World Bank. *Wolfensohn Cancer of Corruption Speech*. https://www.worldbank.org/en/news/video/2022/08/12/wolfensohn-cancer-of-corruption, on 20 Mar. 2022; Diagne, M. Two Decades on, the World Bank Group Remains Committed to our Fight Against Corruption. *World Bank Blog*, 9 Dec. 2020. https://blogs.worldbank.org/governance/two-decades-world-bank-group-remains-committed-our-fight-against-corruption, on 20 Mar. 2022.

[161] TI (2004). *World Bank Move to Reduce Private Sector Bribery Welcomed by Transparency International.* https://www.transparency.org/en/press/world-bank-move-to-reduce-private-sector-bribery-welcomed-by-transparency-i, on 22 Dec. 2023.

[162] AFA. (2023). *Presentation of Various Regulatory Frameworks for Promoting Business Integrity Across the World. https://www.agence-francaise-anticorruption.gouv.fr/files/2023-05/AFA%27s%20Presentation%20FR%20UK%20US%20WBG%20Standards_May%202023_English%20version.pdf.*

[163] OECD, UNODC, World Bank. (2013). *An Anti-Corruption Ethics and Compliance Programme for Business: A Practical Guide*. https://www.oecd.org/corruption/Anti-CorruptionEthicsComplianceHandbook.pdf. To Arlen, "The World Bank sanctions regime allows the bank to exclude individuals and entities for a variety of violations—including both paying bribes and committing fraud. Rather than relying on external authorities to determine whether actionable misconduct occurred – such as local authorities in the recipient country or authorities with jurisdiction over an entity involved in corrupting or defrauding the recipient country – the World Bank regime empowers its own officials to identify and investigate misconduct *sua sponte*." (Jennifer Arlen, *Foreword*, in FIGHTING FRAUD AND CORRUPTION AT THE WORLD BANK: A CRITICAL ANALYSIS OF THE SANCTIONS SYSTEM (Stefano Manacorda & Constantino Grasso, 2018, at xi).

[164] Diagne, M. Two Decades on, the World Bank Group Remains Committed in our Fight Against Corruption. *World Bank Blog*, 9 Dec. 2020. https://blogs.worldbank.org/governance/two-decades-world-bank-group-remains-committed-our-fight-against-corruption, on 20 Mar. 2022.

[165] In addition to investigation and punishment, the INT aims to "create and maintain a trust-based, inclusive organizational culture that encourages ethical conduct, a commitment to compliance with the law and a culture in which Misconduct is not tolerated." (World Bank. *Integrity Vice Presidency*. https://www.worldbank.org/en/about/unit/integrity-vice-presidency, 20 Feb. 2022).

firms or individuals from involvement in projects funded by the World Bank Group.[166]

In 2010, the INT introduced an incentive mechanism for compliance programs within the World Bank Group's scope of operations, by changing the institution's sanction standard: the Integrity Compliance Guidelines.[167] Under this approach, the sanctioned party is no longer automatically released after fulfilling some sanctions.[168] Often, sanctions additionally require the sanctioned entity to implement remedial measures, including, for example, the development and demonstrated implementation of a compliance program.[169]

b. Agreement for Mutual Enforcement of Debarment Decisions (2010)

In 2010, the World Bank Group and the other four leading multilateral development banks (African Development Bank Group, the Asian Development Bank, the European Bank for Reconstruction and Development, and the Inter-American Development Bank Group) signed an agreement providing for mutual and reciprocal enforcement of debarment decisions.[170] The agreement does not mention compliance programs.[171] However, the convergence among the multilateral development banks and, in certain respects, towards a broader convergence among a larger group of international actors,[172] can lead companies doing business with these banks to adopt compliance programs both as a preventative measure and, if wrongful actions have already taken place, as a means of possible mitigation of the severity of sanctions.[173]

c. World Bank Sanctioning Guidelines (2011)

The World Bank Group also offers mitigation incentives for sanctions through its voluntary disclosure program, which rewards a company's cooperation and remedial

---

166 AFA. (2023). *Presentation of Various Regulatory Frameworks for Promoting Business Integrity Across the World.* https://www.agence-francaise-anticorruption.gouv.fr/files/2023-05/AFA%27s%20Presentation%20FR%20UK%20US%20WBG%20Standards_May%202023_English%20version.pdf.

167 World Bank. (2010). *Summary of World Bank Group Integrity Compliance Guidelines*. https://thedocs.worldbank.org/en/doc/06476894a15cd4d6115605e0a8903f4c-0090012011/original/Summary-of-WBG-Integrity-Compliance-Guidelines.pdf.

168 World Bank. (2020). *Integrity Compliance at the World Bank Group: Frequently Asked Questions*. https://thedocs.worldbank.org/en/doc/da26092a692560030e0f2dd5c0a8c07b-0090012020/original/ICO-FAQs-4-2020.pdf.

169 AFA. (2023). *Presentation of Various Regulatory Frameworks for Promoting Business Integrity Across the World.* https://www.agence-francaise-anticorruption.gouv.fr/files/2023-05/AFA%27s%20Presentation%20FR%20UK%20US%20WBG%20Standards_May%202023_English%20version.pdf.

170 Norbert Seiler & Jelena Madir, *Fight against Corruption: Sanctions Regimes of Multilateral Development Banks*, 15 JOURNAL OF INTERNATIONAL ECONOMIC LAW 28 (2012).

171 Asian Development Bank. (2010). *Agreement for Mutual Enforcement of Debarment Decisions.* https://www.adb.org/documents/agreement-mutual-enforcement-debarment-decisions, on 22 Feb. 2024.

172 Frank A. Fariello Jr. & Conrad C. Daly, *Coordinating the Fight against Corruption among MDBS: The Past, Present, and Future of Sanctions*, 45 GEORGE WASHINGTON INTERNATIONAL LAW REVIEW 270 (2013).

173 Norbert Seiler & Jelena Madir, *Fight against Corruption: Sanctions Regimes of Multilateral Development Banks*, 15 JOURNAL OF INTERNATIONAL ECONOMIC LAW 28 (2012).

actions.[174] The World Bank Sanctioning Guidelines prescribe a number of mitigating factors that the relevant decision-makers should consider, including the establishment or improvement of an effective compliance program, which reflects genuine remorse and intention to reform, or may be seen as a calculated step to reduce the severity of the sentence.[175] The Guidelines also provide recommendations on how compliance programs can be imposed or used as a mitigation factor in World Bank Group sanctions. For instance, there is the sanction of debarment with conditional release, where the imposed conditions may include the implementation or improvement of a compliance program.[176]

d. Anti-Corruption Ethics and Compliance Handbook for Business (2013)

In 2013, the World Bank, together with the OECD and the UNODC, published the Anti-Corruption Ethics and Compliance Handbook for Business, aiming to provide a useful resource for companies based in G20 countries and around the world to implement compliance programs.[177] The Handbook compiles international conventions and related documents on the subject produced by various international actors, guiding companies to the best existing practices. The cooperation among international actors in promoting the anti-corruption agenda, and specifically the compliance programs, becomes even clearer when we observe a collective document like this one.

D. International Private Initiatives

1. ICC

The International Chamber of Commerce (ICC) serves as the institutional representative of 45 million companies across more than 170 countries.[178] ICC seeks to promote world trade and investment based on free and fair competition, harmonizes trade practices and formulates terminology and guidelines for importers and exporters,

---

[174] Humboldt-Viadrina School of Governance. (2013). *Motivating Business to Counter Corruption: A Practitioner Handbook on Anti-Corruption Incentives and Sanctions*. https://www.globalcompact.de/migrated_files/wAssets/docs/Korruptionspraevention/Publikationen/motivating_business_to_counter_corruption.pdf.

[175] World Bank. (2011). *World Bank Sanctioning Guidelines.* https://www.worldbank.org/content/dam/documents/sanctions/other-documents/osd/World%20Bank%20Group%20Sanctioning%20Guidelines%20January%202011.pdf. For more information, *see* Norbert Seiler & Jelena Madir, *Fight against Corruption: Sanctions Regimes of Multilateral Development Banks*, 15 JOURNAL OF INTERNATIONAL ECONOMIC LAW 28 (2012).

[176] World Bank. (2011). *World Bank Sanctioning Guidelines.* https://www.worldbank.org/content/dam/documents/sanctions/other-documents/osd/World%20Bank%20Group%20Sanctioning%20Guidelines%20January%202011.pdf; Norbert Seiler & Jelena Madir, *Fight against Corruption: Sanctions Regimes of Multilateral Development Banks*, 15 JOURNAL OF INTERNATIONAL ECONOMIC LAW 28 (2012).

[177] OECD, UNODC, World Bank. (2013). *Anti-Corruption Ethics and Compliance Handbook for Business*. https://web-archive.oecd.org/2019-10-21/256329-Anti-CorruptionEthicsComplianceHandbook.pdf.

[178] ICC. *Our Mission, History and Values.* https://iccwbo.org/about-icc-2/our-mission-history-and-values/.

and provides a range of practical services to business.[179]

a. ICC Rules Against Corruption (1977, 1996, 1999, 2005, 2011)

i. Recommendations to Combat Extortion and Bribery in Business Transactions (1977)

This instrument was produced in 1977 by the Commission on Ethical Practices, an *ad hoc* commission established by the ICC in 1975.[180] This occurred in the context of the repercussions of the global bribery scandals that took place in the 1970s.[181] The Commission, composed of individuals from both developed and developing countries holding high positions in businesses and governments, conducted a survey to assess the existence of legislation prohibiting extortion and bribery worldwide.[182] It concluded that while such regulations exist in most countries, the effectiveness of their enforcement varies considerably.[183] The Commission released the 1977 guideline to address this issue, advocating for complementary and mutually reinforcing actions by states, businesses, and intergovernmental bodies.[184] It was the ICC who "first realised the importance of intergovernmental cooperation in combating international corruption"[185] and "the first business organization to issue anti-corruption rules."[186]

In addition to the foreword, the document has two parts: one directed at governments and the other at businesses. Concerning governments, the 2017 Recommendations advise states to prevent bribery and extortion through several measures. Regarding business, it targets individuals or entities engaged in business to promote self-regulation in the international business arena in a section named Rules of Conduct to Combat Extortion and Bribery. The Rules are crafted as a voluntary framework applicable to enterprises of all sizes and in all countries, outlining five basic rules and six guidelines for their implementation. This aims to help companies establish

---

[179] Antonio Argandoña, *The 1996 ICC Report on Extortion and Bribery in International Business Transactions,* 6 BUSINESS ETHICS, THE ENVIRONMENT AND RESPONSIBILITY 134 (1997). Founded in 1919 in the aftermath of World War I, the International Chamber of Commerce (ICC) was established as a response to the absence of a global framework concerning governing trade, investment, finance, and commercial relations. It was founded by a group of industrialists, financiers, and traders who referred to themselves as the "Merchants of Peace" and believed that the private sector was best suited to establish global business standards. (ICC. *Our Mission, History and Values.* https://iccwbo.org/about-icc-2/our-mission-history-and-values/, on 5 Jun. 2023).

[180] ICC, *Commission on Ethical Practices Recommendations to Combat Extortion and Bribery in Business Transactions*, 17 INTERNATIONAL LEGAL MATERIALS 417 (1978). http://www.jstor.org/stable/20691864, at 418.

[181] ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf.

[182] ICC, *Commission on Ethical Practices Recommendations to Combat Extortion and Bribery in Business Transactions*, 17 INTERNATIONAL LEGAL MATERIALS 417 (1978). http://www.jstor.org/stable/20691864.

[183] ICC, *Commission on Ethical Practices Recommendations to Combat Extortion and Bribery in Business Transactions*, 17 INTERNATIONAL LEGAL MATERIALS 417 (1978). http://www.jstor.org/stable/20691864.

[184] ICC, *Commission on Ethical Practices Recommendations to Combat Extortion and Bribery in Business Transactions*, 17 INTERNATIONAL LEGAL MATERIALS 417 (1978). http://www.jstor.org/stable/20691864.

[185] Joseph Mase, *Fighting Transnational Corruption*, 9 AMICUS CURIAE 4 (1998), at 4.

[186] ICC. (2011). *ICC Rules on Combating Corruption*. https://iccwbo.org/wp-content/uploads/sites/3/2011/10/ICC-Rules-on-Combating-Corruption-2011.pdf, at 3.

effective control systems to prevent extortion and bribery, focusing on the process of obtaining and retaining business with the public or private sector. The document does not expressly mention "compliance programs."

ii. Revisions to the ICC Rules of Conduct on Extortion and Bribery in International Business Transactions (1996)

During the 1990s, a new wave of corruption scandals emerged, reigniting global attention toward responding to it.[187] In 1994, the ICC established an *ad hoc* committee to review the 1977 Recommendations.[188] The 1996 document emphasized that the 1977 Recommendations generated interest in intergovernmental fora, such as the OECD and the UN, and motivated corporations in various countries to establish or strengthen their internal rules of fair practices, using the Rules of Conduct as a guide. The ICC asserts that, at the time, virtually all countries prohibit extortion and bribery, unlike in 1977.

The 1996 document expands the 1977 recommendations for governments to include international organizations, underscoring the significance of these institutions in the global anti-corruption effort. However, the content of the guidelines remains largely the same. The most notable modifications in the 1996 document were the ones targeting companies. Although brief compared to the 1977 ones, the 1996 Rules for companies impose more rigorous measures by encompassing bribery in all aspects, beyond merely the acquisition and preservation of business as the 1997 version. The document also recommends that states, whether members or non-members of the OECD, adopt the 1994 OECD Recommendation on Bribery in International Business Transactions.[189] Moreover, the document demands actions against corruption from international financial institutions, namely the World Bank, which the ICC understands should take reasonable steps to ensure that corrupt practices do not occur in connection with projects they finance. Additionally, the document calls for more involvement of the WTO in the fight against corruption. The 1996 documents also do not expressly mention compliance programs, although they make recommendations for companies connected to a compliance program.

iii. ICC Rules of Conduct on Extortion and Bribery in International Business Transactions (1999)

The 1999 edition essentially reissues the 1996 guidelines, incorporating minor modifications.[190] Among the additions, the document highlights the development of

---

[187] ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf.

[188] ICC. *1996 Revisions to the ICC Rules of Conduct on Extortion and Bribery in International Business Transactions*, 35 INTERNATIONAL LEGAL MATERIALS, 1306 (1996). http://www.jstor.org/stable/20698610.

[189] OECD, *Council Recommendation on Bribery in International Business Transactions*, 33 INTERNATIONAL LEGAL MATERIALS 1389 (1994). http://www.jstor.org/stable/20698384.

[190] ICC. (1999). *ICC Rules of Conduct: Extortion and Bribery in International Business Transactions – 1999 revised version*. https://1997-2001.state.gov/global/narcotics_law/global_forum/F810bocr.pdf.

more anti-bribery initiatives around the world.

iv. Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations (2005)

In this document, the ICC stated that due to the increasing progress in anti-corruption efforts, highlighting the UN Convention, the ICC Commission on Anti-Corruption decided to "revisit and rethink the ICC Rules of Conduct and to refine its stance on a number of integrity matters."[191] This was the first time that the document was not produced by an ad hoc Commission. The 2005 version of the ICC Rules brought updates compared to the previous editions, maintaining the core principles and substance.

Regarding compliance programs, although they were not explicitly mentioned in the Rules, the introduction to the document affirms that the success of the ICC Rules depends on a clear message from the company's chief executive that bribery and extortion are prohibited and that an effective compliance program will be implemented. Furthermore, in the section targeting government, it suggests that governments make the adoption of anti-corruption compliance programs a condition for major government contracts. Thus, the 2005 edition is the first that expressly mentions compliance programs.

In the section intended for international organizations and governments, the ICC acknowledges the widespread recognition and progress made in combating corruption, particularly in strengthening legal frameworks around the world. The ICC emphasizes its endorsement of the OECD and UN Conventions, as well as other regional agreements, while stressing the need for greater coordination to address the existing inconsistency and lack of common definitions from an international business perspective. Moreover, the ICC commends initiatives from institutions such as the World Bank and IMF, urging them to go even further by incorporating requirements for contractors to adopt anti-bribery compliance programs. Similarly, the ICC encourages the Global Compact Office to promote the adoption of corporate compliance programs consistent with ICC Rules among companies participating in the Global Compact. This reveals that the ICC was a pioneer among the actors of the IACR in promoting compliance programs and also urged other actors to do the same.

v. ICC Rules on Combating Corruption (2011)

The 2011 document reflects major changes compared to previous versions.[192] The 2011 version has no specific provisions on governments or international organizations, declaring the focus to be a non-binding method of self-regulation for businesses, in light of the international legal instruments.[193] The 2011 edition has three

---

[191] ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf, at 3.

[192] ICC. (2011). *ICC Rules on Combating Corruption.* https://iccwbo.org/wp-content/uploads/sites/3/2011/10/ICC-Rules-on-Combating-Corruption-2011.pdf. This Commission affirms be a leading body in the development of rules of conduct, best practices, and advocacy for fighting corruption and for corporate responsibility, working closely with intergovernmental organizations, such as the UN and the OECD.

[193] The document explicitly lists these instruments. *Global Instruments*: United Nations Convention Against Corruption (UNCAC); United Nations Convention Against Transnational Organized Crime (UNTOC); OECD Convention on the Bribery of Foreign Public Officials in International Business

parts: (i) the ICC Rules on Combating Corruption; (ii) policies that companies should enact to support compliance with the Rules; and (iii) a list of the suggested elements of an effective corporate compliance program.[194] For the ICC, the 2011 version "mirror-images the impressive evolution of the ethics and compliance practices of leading enterprises."[195]

In this version, for the first time, compliance programs are included in the Rules, which outline the elements for the effectiveness of this strategy. The document also encourages collective action, such as anti-corruption pacts related to specific projects or long-term anti-corruption initiatives involving the public sector and/or peers in the respective business sectors. The preface and introduction of the document emphasize that small and medium-sized enterprises should also adopt compliance programs.

b. ICC Handbook (1999, 2003, 2008)

The summary of the 2005 ICC Rules highlights the publication by the ICC of the manual named Fighting Corruption: A Corporate Practices Manual in 1999 and extensively revised and republished in 2003, providing "detailed practical guidance for compliance with the ICC Rules of Conduct and the OECD."[196] The analysis of these manuals was not included in this article due to their lack of accessibility, as the ICC has not made them more available. In 2008, the document was once again updated and published under the name Fighting Corruption: International Corporate Integrity Handbook, which is available on the ICC page and, due to this, I analyzed it in this article.[197]

Transactions (OECD Convention); OECD Recommendation for Further Combating Bribery of Foreign Public Officials in International Business Transactions, including Annex II Good Practice Guidance on Internal Controls, Ethics and Compliance. *Africa*: African Union Convention on Preventing and Combating Corruption (AU Convention); Southern African Development Community Protocol Against Corruption (SADC Protocol); Economic Community of West African States Protocol on the Fight Against Corruption (ECOWAS Protocol). *Americas*: Inter-American Convention Against Corruption (OAS Convention). *Asia and Pacific region*: ADB-OECD Action Plan for Asia-Pacific (Action Plan). *Europe*: Council of Europe Criminal Law Convention; Council of Europe Civil Law Convention; Resolution of the Committee of Ministers of the Council of Europe: Agreement Establishing the Group of States Against Corruption; Resolution of the Committee of Ministers of the Council of Europe: Twenty Guiding Principles for the Fight Against Corruption; European Union Convention on the Protection of the Communities' Financial Interests and the Fight Against Corruption and two related Protocols; European Union Convention on the Fight Against Corruption involving officials of the European Communities or officials of Member States. *See,* Appendix A (ICC. (2011). *ICC Rules on Combating Corruption.* https://iccwbo.org/wp-content/uploads/sites/3/2011/10/ICC-Rules-on-Combating-Corruption-2011.pdf, at 13).

[194] ICC. *ICC Rules on Combating Corruption.* https://iccwbo.org/news-publications/policies-reports/icc-rules-on-combating-corruption/

[195] ICC. (2011). *ICC Rules on Combating Corruption.* https://iccwbo.org/wp-content/uploads/sites/3/2011/10/ICC-Rules-on-Combating-Corruption-2011.pdf, at 3.

[196] ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf, at 3.

[197] The only version available on the ICC site is the 2008 version. *See*, ICC. *Fighting Corruption – International Corporate Integrity Handbook*. https://2go.iccwbo.org/fighting-corruption.html, on 11 Aug. 2023.

The book, among other reflections, looks ahead at priorities in the fight against corruption, including reflections on compliance programs.[198] It asserts that adopting ethical principles is easy, but formulating detailed compliance programs and integrating them into corporate culture are harder, highlighting that external verification of the program remains controversial. The Handbook concludes by emphasizing the importance of overcoming obstacles in the fight against corruption, given its detrimental impact on the global economy, democratic institutions, and international development, as well as in the obstacles in implementation of compliance programs.

c. Other ICC Initiatives

The ICC has been developing anti-corruption tools and specific guidelines on compliance programs elements to help companies implement the ICC Rules.[199] The ICC also participated in a joint publication with UN Global Compact, TI, and PACI, named *Clean Business is Good Business*.[200] It offers a summary of arguments and information to help companies make the business case against corruption, including implementing anti-corruption programs.[201]

2. TI

WRC[202] pointed the TI as part of the IACR.[203] TI defines itself as a global movement working in over 100 countries with the mission to stop corruption and promote transparency, accountability and integrity at all levels and across all sectors of society.[204] TI was founded in 1993 when corruption was a taboo topic.[205] Witnessing the impact of corruption during his work in East Africa, retired World Bank official Peter Eigen and nine allies established a small organization to address this taboo, which later became TI.[206]

---

[198] Fritz Heimann & Mark Pieth, *Moving Anti-corruption to the Next Level*, in FIGHTING CORRUPTION: INTERNATIONAL CORPORATE INTEGRITY HANDBOOK 209 (Fritz Heimann & François Vincke ed., 2008).

[199] The documents concerning elements could be part of compliance programs and not about compliance programs themselves; thus, I did not analyze them in this article. For instance, they published the ICC Guidelines on Gifts and Hospitality, the ICC Anti-Corruption Third Party Due Diligence: A Guide for Small and Medium Size Enterprises, and the ICC Guidelines on Whistleblowing. *See*, ICC. *ICC Rules on Combating Corruption.* https://iccwbo.org/news-publications/policies-reports/icc-rules-on-combating-corruption/.

[200] ICC, TI, UN Global Compact, and PACI. (2008). *Clean Business is Good Business*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2Fclean_business_is_good_business.pdf.

[201] The ICC has also been developing anti-corruption tools and specific guidelines on compliance programs elements to help companies implement the ICC Rules. For instance, they published the ICC Guidelines on Gifts and Hospitality, the ICC Anti-Corruption Third Party Due Diligence: A Guide for Small and Medium Size Enterprises, and the ICC Guidelines on Whistleblowing.

[202] Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013)

[203] "The Corruption Perceptions Index (CPI) is the most widely used global corruption ranking in the world. It measures how corrupt each country's public sector is perceived to be, according to experts and businesspeople." (TI. *The ABCs of the CPI: How the Corruption Perceptions Index is Calculated*. https://www.transparency.org/en/news/how-cpi-scores-are-calculated, on 5 Dec. 2023).

[204] TI. *About*. https://www.transparency.org/en/about, on 5 Dec. 2023.

[205] TI. *Our Story*. https://www.transparency.org/en/our-story, on 5 Dec. 2023.

[206] TI. *Our Story*. https://www.transparency.org/en/our-story, on 5 Dec. 2023.

The PACI mentions that the origins of their Principles, which promote compliance programs, lie in the TI Business Principles for Countering Bribery from 2002.[207] Furthermore, the Business Against Corruption: A Framework for Action – published by UN, TI, and International Business Leaders Forum in 2005 –, among other documents here analyzed, also references the TI Business Principles for Countering Bribery.[208] Therefore, the analysis of TI's push for compliance programs will start with this pioneering document and its updates.

a. Business Principles for Countering Bribery (2002, 2003, 2004, 2008, 2009, 2013, 2015)

i. Business Principles for Countering Bribery: An Initiative of Transparency International and Social Accountability International (2002)[209]

The Business Principles initiative began in 1999 when TI and its partners recognized the potential to complement the OECD Convention.[210] Feasibility study was conducted, leading to the formation of a Steering Committee composed of representatives from both business and civil society.[211] Their collaborative effort aimed to determine if a consensus framework could be developed for private sector

[207] WEF. (2004). *Partnering Against Corruption – Principles for Countering Bribery.* https://media.corporate-ir.net/media_files/irol/70/70435/PACI.pdf.

[208] UN Global Compact, TI, International Business Leaders Forum. (2005). *Business Against Corruption: A Framework for Action – Implementation of the 10th UN Global Compact Principle Against Corruption*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2F7.7%2FBACtextcoversmallFINAL.pdf.

[209] TI and Social Accountability International. (2002). *Business Principles for Countering Bribery: An Initiative of Transparency International and Social Accountability International.* https://www.newsd.admin.ch/newsd/message/attachments/5465.pdf. This document was located through a Google search using the same term "TI Business Principles for Countering Bribery 2002" on December 6, 2023. It was not listed in the official TI page results when searching for "Business Principles for Countering Bribery" using the site's search engine. This search, on December 6, 2023, produced 42 results, none of which is this document. On the same day, a search on TI's official page for this term, applying the "publication" filter (excluding news, blog, country profile, priority, advocacy, and events), produced four results. These results include the following documents, in this order: (i) Business Principles for Countering Bribery (2013); (ii) Business Principles for Countering Bribery: Small and Medium Enterprise (SME) Edition; (iii) Assurance Framework for Corporate Anti-bribery Programs (2012); (iv) Anti-Corruption Principles for State-Owned Enterprises: A Multi-Stakeholder Initiative of Transparency International (2017). The first three will be discussed here as they are relevant to this article. The rest of the TI documents in this section I also found through a free search on Google, as they did not appear in the TI's site search with or without filters.

[210] TI. (2004). *Business Principles for Countering Bribery: Guidance Document.* https://www.ethics.org/wp-content/uploads/resources/Business_Principles_for_Countering_Bribery_Transparency_Intl_Guidance_Document_2004.pdf.

[211] TI. (2004). *Business Principles for Countering Bribery: Guidance Document.* https://www.ethics.org/wp-content/uploads/resources/Business_Principles_for_Countering_Bribery_Transparency_Intl_Guidance_Document_2004.pdf.

use.[212] The Business Principles for Countering Bribery are a result of this effort.[213] The Principles are: (i) the enterprise shall prohibit bribery in any form, whether direct or indirect; (ii) the enterprise shall commit to the implementation of a program to counter bribery. This document also provides recommendations about the development of a compliance program, asserting the aim to addresses the need for companies to respond to increasing regulatory demands and heightened awareness of bribery risks. This document is a reissue of the 2003 document, with no alterations to the text. [214]

ii. Business Principles for Countering Bribery: Guidance Document (2004)

This guidance presents the same Principles as well as the same recommendations on the scope and requirements of the compliance programs as the 2002 document.[215] However, it has the broader aim to provide background and clarification to the 2002 Business Principles. Another difference is that the 2004 document has more details on how enterprises can create and implement the anti-bribery program. For example, it outlines a Six-Step Process for program implementation.[216]

Among the new information, TI asserts that the Principles are considered good practices, not best practices, as it is expected that they will further strengthen over time. In this sense, the document starts saying that it does not constitute, and does not purport to constitute, definitive statements of TI policies in anti-bribery, being inappropriate, at the time, to try to establish definitive policies in such a rapidly developing area. In addition, it makes clear that the Business Principles focus on bribery and not corruption in general.

The document highlights that an anti-bribery program is different from the traditional corporate compliance function, as it is usually understood as only ensuring observation of legal requirements. The documents understands that a legally based compliance can quickly become unmanageable for international companies because of the different laws in operation in each country. Thus, its advocacy that companies should adopt an anti-bribery program, which is embed a culture of avoiding bribery into their business functions, doing more than just comply with domestic rules.

---

[212] TI. (2004). *Business Principles for Countering Bribery: Guidance Document.* https://www.ethics.org/wp-content/uploads/resources/Business_Principles_for_Countering_Bribery_Transparency_Intl_Guidance_Document_2004.pdf.

[213] TI and Social Accountability International. (2002). *Business Principles for Countering Bribery: An Initiative of Transparency International and Social Accountability International.* https://www.newsd.admin.ch/newsd/message/attachments/5465.pdf.

[214] TI. (2003). *Business Principles for Countering Bribery: An Essential Tool.* https://www.mohw.gov.tw/dl-15587-7a3701c0-05e5-4d74-bb9f-57a9390b2c58.html.

[215] TI. (2004). *Business Principles for Countering Bribery: Guidance Document.* https://www.ethics.org/wp-content/uploads/resources/Business_Principles_for_Countering_Bribery_Transparency_Intl_Guidance_Document_2004.pdf.

[216] In 2005, a document regarding these steps was separately published under the name TI Six-Step Process: A Practical Guide for Companies Implementing Anti-Bribery Policies and Programmes, according to Hess (David Hess, *Partnering Against Corruption Initiative and the Business Principles for Countering Bribery*, in HANDBOOK OF TRANSNATIONAL GOVERNANCE INSTITUTION & INNOVATIONS 322 (Thomas Hale & David Held ed., 2011). However, this TI document was not located.

The guidance also declares to be stimulated vis-a-vis the new developments on the subject: the UN Convention and the introduction of the UN Global Compact 10th Principle against Corruption, which will take some time to be enforced in all signatory countries. Recognizing that companies need tools to help them break the cycle of corruption, TI affirms that the Principles, for the first time, provide a comprehensive approach to countering bribery by companies. The document also declares that the TI Principles have been evaluated widely through field-tests and workshops since the first publication and have been endorsed or adopted by leading multinationals. It affirms that the international demand – in both North and South Global and from different industry groups – shows that a tool to help companies implement a no-bribes policy is really needed.

The document states that, up until its publication, there had been a growing emphasis on requirements for the private sector in terms of business ethics and the broader concept of corporate responsibility. It asserts that, in a market economy, the short-term focus on maximizing returns to shareholders needs to be replaced by a longer-term orientation to the demands of stakeholders as a prerequisite for corporate sustainability. This includes the fight against corruption, as corrupt business practices pose a serious risk to the long-term sustainability of businesses and can significantly undermine reputation and shareholder value. For that reason, it affirms that the Business Principles are designed to strike a balance between a values-based approach and a compliance-based approach.

Regarding the company's anti-bribery measures, the document affirms that bribery can take place through agents and intermediaries, and to avoid this, it suggests that companies adopt integrity pacts. The document makes it clear that an integrity pact – a form of collective action cited in some documents analyzed in this article – is a tool developed in the 1990s by TI to help governments, businesses, and civil society in the fight against corruption in the field of public contracting. However, it is applicable across all contracting to guarantee that subsidiaries, joint venture partners, agents, contractors, and other third parties with the company have business relationships also do not engage in bribery.

iii. Business Principles for Countering Bribery: Small and Medium Enterprise (SME) Edition (2008)

The document acknowledges that much of the world's business is conducted SMEs, especially in emerging economies.[217] Consequently, the adherence to an anti-bribery system by small companies is crucial to success in the global fight against corruption. It also recognizes that SMEs – which typically have fewer resources in terms of time, money, and employees – face challenges in resisting and countering bribery pressures. In contrast, large international companies had been increasingly requiring their SMEs ant other suppliers to provide evidence of having appropriate anti-bribery policies and systems in place. Thus, the document aims to outline, in a clear and direct manner, the process by which smaller businesses can develop an anti-bribery compliance program aligned to their size and resources. This version presents a simplified process for implementing anti-bribery programs compared to the 2002 guide,

[217] TI. (2008). *Business Principles for Countering Bribery: Small and Medium Enterprise (SME) Edition.* https://www.transparency.org/en/publications/business-principles-for-countering-bribery-small-and-medium-enterprise-sme.

clarifying potential issues and offering practical examples, guided by the Business Principles.

iv. Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International (2009)

The 2009 edition represents a light revision of the 2002 Business Principles, aimed at accommodating developments in key areas of good practice and aligning, where appropriate, with other leading anti-bribery codes such as the ICC Rules and the PACI Principles.[218] The document emphasizes that the value of the Principles has been proven through consultations, field testing, and workshops. It also states that although surveys indicate that companies are adopting anti-bribery policies, full implementation remains an incomplete process and a challenge for many. In this version, TI hopes that companies will increasingly utilize the Business Principles, leading to a higher and more uniform standard of anti-bribery practice worldwide, thus contributing to a more level playing field.

TI reinforces in the document that an effective anti-bribery program not only strengthens reputation but also builds the respect of employees, enhances credibility with key stakeholders, and supports an enterprise's commitment to corporate responsibility. Notably, the 2009 version introduces a new section on external verification and assurance, suggesting that the companies' board should consider commissioning external verification or assurance to enhance internal and external confidence in the program's effectiveness.

v. Business Principles for Countering Bribery: Transparency International Self-Evaluation Tool (2009)

TI's Self-Evaluation Tool (SET) is a checklist that allows companies to assess their anti-bribery programs aligned with the 2009 Business Principles for Countering Bribery.[219] With a focus on a zero-tolerance policy, the SET aims to guide companies in implementing effective anti-bribery measures, ensuring alignment with assessed risks and stakeholder confidence. The tool provides indicators that can be used for external reporting, internal performance metrics, and supports internal audit.

---

[218] TI. (2009). *Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International.* https://www.pactomundial.org/wp-content/uploads/2015/04/Principios-Empresariales-para-Contrarrestar-el-Soborno-de-Transparencia-Internacional-Inglés.pdf.

[219] TI. (2009). *Business Principles for Countering Bribery: Transparency International Self-Evaluation Tool.* https://www.transparency.org/files/content/tool/2009_TI_BusinessSelfEvaluationTool_EN.pdf.

vi. Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International (2013)[220]

This document is the second revision of the 2002 Business Principles, encompassing a broader scope the 2009 revision.[221] In this version, TI asserts that significant changes have occurred since the initial publication. The landscape has evolved markedly with the introduction of more stringent domestic and foreign bribery laws, heightened enforcement, substantial fines, and the looming prospect of sanctions for company directors and employees, all of which have reverberated throughout the business community. Moreover, TI affirms that mounting pressures from socially responsible investment funds and indices, incorporating anti-bribery criteria into their screening procedures, contribute to the evolving landscape.

The document declares its understanding of these recent developments in anti-bribery practices and, consequently, incorporates modifications to the original text. The adjustment aims to underscore the contemporary significance of these issues in anti-bribery practices and the feedback that TI received since the last update, fostering closer alignment with other leading codes and legal instruments, notably the UN Convention. Among the changes, an additional part was incorporated into the second principle of the 2002 version: "The Programme shall represent the enterprise's anti-bribery efforts including values, code of conduct, detailed policies and procedures, risk management, internal and external communication, training and guidance, internal controls, oversight, monitoring and assurance," aiming to stimulate more robust programs.

vii. Business Principles for Countering Bribery: Commentary (2015)

This document is a commentary on the 2013 edition of the Business Principles, explaining the 2013 edition's changes, providing background to its provisions, and offering insights into implementation.[222] For instance, the document asserts that some changes aimed to stimulate companies in fostering a culture of integrity within the enterprise through the compliance program.

The document also emphasizes that the 2013 Business Principles, the version current in force, aim to serve as best practice guidance for enterprises countering bribery, influencing corporate anti-bribery practices, and serving as a reference for various frameworks, both domestically and internationally. This reflects a maturity of the Principles, unlike the 2004 version which stated that it was inappropriate, at the time, to try to establish best practices in the area. Moreover, the Commentary affirms the Business Principles' influence as a business benchmark, fostering the development and strengthening of anti-bribery measures for enterprises globally. The document also

---

[220] TI. (2013). *Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International.* https://www.transparency.org/en/publications/business-principles-for-countering-bribery.

[221] TI. (2013). *Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International.* https://www.transparency.org/en/publications/business-principles-for-countering-bribery.

[222] TI. (2015). *Business Principles for Countering Bribery: Commentary.* https://www.transparency.org/files/content/publication/2015_BusinessPrinciplesCommentary_EN.pdf.

justifies maintaining the focus on bribery, highlighting the significant impact of this misconduct on enterprises and societies.

b. Other TI initiatives

i. Assurance Framework for Corporate Anti-bribery Programs (2012)

Supported by the WEF, TI affirms to have developed this instrument in response to a grew demand for comprehensive and continuously monitored anti-bribery initiatives in business dealings.[223] This voluntary framework aims to standardize the design of robust anti-bribery programs, providing an assurance process for companies to assess and enhance the strength and credibility of their initiatives. The document asserts that it addresses the rising expectation for enterprises to transparently communicate their anti-bribery measures to stakeholders, bridging the credibility gap created by corporate bribery scandals and skepticism among stakeholders regarding anti-bribery efforts. The Assurance Framework is part of TI's toolkit based on the Business Principles, consisting of five stages and objectives covering the environment, risk assessment, control activities, information and communication, and monitoring.

ii. Business Integrity Programme Project

This project seeks address the global challenge of corruption in various sectors, emphasizing collaboration with businesses, governments, and civil society. TI informs that the program engages in multi-stakeholder partnerships, advocating for a strong anti-corruption environment, promoting ethical business practices, and fostering anti-corruption culture. It operates through thematic projects, involving selected businesses in specific areas like business purpose, professional services, technology, and integrity tools. It also aims to facilitates multi-stakeholder collaboration through expert guidance councils and business integrity boards. One of the publications related to this project is Stories of Change: Better Business by Preventing Corruption.[224] It highlights the benefits of strong compliance programs, such as improved business performance, promotion of fair competition, minimization of losses due to corruption, increased access to capital, and enhanced reputations.[225] This project seems to have a broader scope, encompassing corruption beyond just bribe.

---

[223] TI. (2012). *Assurance Framework for Corporate Anti-bribery Programs.* https://transparency.org.au/wp-content/uploads/2020/09/Report_Corporate-Antibribery.pdf.

[224] TI. (2018). *Stories of Change: Better Business by Preventing Corruption*. https://images.transparencycdn.org/images/2018_Report_StoriesOfChange_English.pdf.

[225] The document describes four case studies: the TI Business Integrity Programme: Transparency in Corporate Reporting (TRAC), focusing on driving disclosure to prevent corruption through a report series; the Indonesia case involving partnerships with Perusahaan Listrik Negara, the state-owned electricity supplier, to enhance transparency; the Italy case, where TI collaborates with large businesses to promote anti-corruption practices down the supply chain using an integrity kit for small- and medium-sized enterprises; and lastly, the Mexico case, where TI has partnered with businesses and civil society organizations to advocate for robust anti-corruption laws regulating the country's business sector. (TI. (2018). *Stories of Change: Better Business by Preventing Corruption*. https://images.transparencycdn.org/images/2018_Report_StoriesOfChange_English.pdf).

3. WEF PACI

WRC pointed PACI, an initiative from the WEF,[226] as part of the IACR.[227] WRC described PACI as a voluntary code of conduct initiative that corporations can choose to join.[228] Presently, 80 organizations from various business sectors globally are participants,[229] including some companies implicated in corruption scandals like Petrobras, which was a Forum member from 2005 to 2014, rejoined in 2020.[230] PACI born in 2004[231] and currently positions itself as a CEO-led platform in the global anti-corruption arena, emphasizing public-private cooperation, responsible leadership, and technological advances.[232] The PACI stands that "fighting corruption in all its forms not only advances the development and well-being of society but also makes businesses stronger, more resilient to risk, more ethical and, ultimately, more sustainable."[233] In summary, it affirms that compliance programs are good for business, as other

---

[226] The WEF, an NGO based in Geneva, is dedicated to demonstrating entrepreneurship in the global public interest while maintaining the highest standards of governance (WEF. *Our Mission*. https://freedomhouse.org/policy-recommendations/combatting-corruption-and-kleptocracy, on 30 Nov. 2023). In 1973, WEF announced the Davos Manifesto, outlining a code of ethics for business leaders. It emphasizes that the purpose of professional management is to serve clients, shareholders, workers, and society, with a commitment to competitiveness, shareholder returns, employee well-being, and societal responsibility, all underpinned by the necessity of profitability for long-term sustainability (WEF. (2019). *Davos Manifesto 1973: A Code of Ethics for Business Leaders*. https://www.weforum.org/agenda/2019/12/davos-manifesto-1973-a-code-of-ethics-for-business-leaders/, on 30 Nov. 2023). WEF published an updated version of the Manifesto, the "Davos Manifesto 2020: The Universal Purpose of a Company in the Fourth Industrial Revolution," which express the WEF vision of "stakeholder capitalism" and affirm a "zero tolerance for corruption." (WEF. (2019). *Davos Manifesto 2020: The Universal Purpose of a Company in the Fourth Industrial Revolution.* https://www.weforum.org/agenda/2019/12/davos-manifesto-2020-the-universal-purpose-of-a-company-in-the-fourth-industrial-revolution/, on 30 Nov. 2023).

[227] Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[228] Jan Wouters, Cedric Ryngaert & Ann Sofie Cloots, *The International Legal Framework Against Corruption: Achievements and Challenges*, 14 MELBOURNE JOURNAL OF INTERNATIONAL LAW 205 (2013).

[229] WEF. *Partnering Against Corruption Initiative – PACI Signatories.* https://www3.weforum.org/docs/Partnering_Against_Corruption_Initiative_Members_2021.pdf, on 30 Nov. 2023.

[230] SEC. (2020). *Form 6-K: Petrobras Returns to the World Economic Forum's Anti-Corruption Initiative.* https://www.sec.gov/Archives/edgar/data/1119639/000129281420003450/pbra20200909_6k.htm#:~:text=Rio%20de%20Janeiro%2C%20September%209,of%20the%20World%20Economic%20Forum, on 30 Nov. 2023.

[231] Originally named "Industry Partnership Programme," launched in 2004, it initially focused on combating bribery in three pilot sectors: IT and telecommunications, energy, and financial services. (WEF. *WEF History: 2004*. https://widgets.weforum.org/history/2004.html, on 30 Nov. 2023). The initiative changed name in 2005, and also adopted a multisector approach (Lee Tashjian, *Partnering Against Corruption Initiative Leads Industry Battle Against Corruption*, 9 LEADERSHIP AND MANAGEMENT IN ENGINEERING 123 (2009)), to address various industry, regional, country, or global anti-corruption issues based on member companies' needs and interests (WEF. *Partnering Against Corruption Initiative.* https://www.weforum.org/communities/partnering-against-corruption-initiative/, on 30 Nov. 2023).

[232] WEF. *Partnering Against Corruption Initiative.* https://www.weforum.org/communities/partnering-against-corruption-initiative/, on 30 Nov. 2023.

[233] WEF. (2016) *Partnering Against Corruption Initiative Global Principles for Countering Corruption: Application & General Terms of Partnership.* https://www3.weforum.org/docs/WEF_PACI_Global_Principles_for_Countering_Corruption.pdf.

documents mentioned here.

a. PACI Principles (2004, 2014)[234]

i. PACI Principles for Countering Bribery (2004)[235]

The 2004 PACI Principles comprised two key elements: (i) a commitment by the enterprise to prohibit bribery in any form and (ii) a commitment to maintaining or implementing an effective anti-bribery program.[236] It also provides practical guidance applicable to companies of all sizes, assisting them in developing policies and programs to combat bribery and corruption in international business. The 2004 Principles declare the aim to give practical effect to the OECD Convention and other governmental and private sector anti-corruption initiatives, such as ICC.[237]

The document affirms that the 2004 Principles were built on the general industry anti-bribery principles developed in 2002 by TI, the Business Principles for Countering Bribery.[238] A key distinction between PACI Principles and TI Principles strategies lies

---

[234] The information on the "Principles for Countering Bribery" and the "Principles for Countering Corruption" is no longer present on the initiative's homepage. In an effort to recover these principles, referenced in other documents analyzed during this article, I utilized the WEF website search mechanism with the query "Principles for Countering" on November 30, 2023. I identified five results, listed in the following order: (i) WEF. (2016) *Partnering Against Corruption Initiative Global – Principles for Countering Corruption: Application & General Terms of Partnership.* https://www3.weforum.org/docs/WEF_PACI_Global_Principles_for_Countering_Corruption.pdf; (ii) WEF. (2014). *World Economic Forum Calls on Business Leaders to Strive for Corruption-Free World.* https://www.weforum.org/press/2014/01/world-economic-forum-calls-on-business-leaders-to-strive-for-corruption-free-world/; (iii) WEF. (2018) *UN Global Compact Communication on Engagement.* https://www3.weforum.org/docs/WEF_UN_Global_Compact_Communication_on_Engagement2018.pdf; (iv) WEF. (2012). *Annual Meeting of the New Champions 2012.* https://www3.weforum.org/docs/AMNC12/WEF_AMNC12_Factsheet.pdf; (v) WEF. (2010). *Everybody's Business: Strengthening International Cooperation in a More Interdependent World: Report of the Global Redesign Initiative.* https://www3.weforum.org/docs/WEF_GRI_EverybodysBusiness_Report_2010.pdf. Only the first two are relevant to this study. Furthermore, I searched for these expressions on Google, on the same date, and located the PACI Principles for Countering Bribery from 2004, which I analyzed in this section.

[235] In January 2004, an initial version of the PACI Principles was developed, focusing specifically on the Engineering and Construction sector. (WEF. (2004). Partnering Against Corruption – Principles for Countering Bribery. https://media.corporate-ir.net/media_files/irol/70/70435/PACI.pdf). This precursor version emerged from a core group of CEOs participating in the "Industry Partnership Programme" concerning about the detrimental impact of corruption on business and society and recognized the need for a coordinated response. (WEF. (2016). Partnering Against Corruption Initiative Global – Principles for Countering Corruption: Application & General Terms of Partnership. https://www3.weforum.org/docs/WEF_PACI_Global_Principles_for_Countering_Corruption.pdf). By October 2004, the document expanded to gain support from companies across various industries, adopting the name "Partnering Against Corruption – Principles for Countering Bribery." (WEF. (2004) *Partnering Against Corruption – Principles for Countering Bribery.* https://media.corporate-ir.net/media_files/irol/70/70435/PACI.pdf). This set of principles resulted from collaboration within a task force composed of member companies from the WEF, TI, and the Basel Institute on Governance (WEF. (2004) *Partnering Against Corruption – Principles for Countering Bribery.* https://media.corporate-ir.net/media_files/irol/70/70435/PACI.pdf).

[236] WEF. (2004) *Partnering Against Corruption – Principles for Countering Bribery.* https://media.corporate-ir.net/media_files/irol/70/70435/PACI.pdf.

[237] WEF. (2004) *Partnering Against Corruption – Principles for Countering Bribery.* https://media.corporate-ir.net/media_files/irol/70/70435/PACI.pdf.

[238] TI and Social Accountability International. (2002). *Business Principles for Countering Bribery: An Initiative of Transparency International and Social Accountability International.* https://www.newsd.admin.ch/newsd/message/attachments/5465.pdf.

in their approach to obtaining anti-corruption commitments.[239] PACI focuses on securing commitments from top management through public declarations and the adoption of principles, with signatories publicly listed on the PACI website.[240] In contrast, the TI approach does not prioritize public adoption of the principles.[241]

ii. PACI Principles for Countering Corruption (2014)[242]

At the Annual PACI Task Force Meeting, in 2013, a review of the PACI Principles was discussed, considering the significant developments in corporate compliance and the emergence of several new compliance and integrity instruments.[243] Following the review, the Principles were signed in 2014 by leading companies that are PACI members, under the name "PACI Principles for Countering Corruption."[244] This update broadens the focus beyond bribery, aligning with the evolving global fight against corruption.[245]

b. Agenda for Business Integrity (2019)[246]

In 2019, the WEF created the Global Future Council on Transparency and Anti-Corruption, comprised of experts, to develop the "Agenda for Business Integrity" that currently guides the PACI's strategy.[247] The Agenda outlines four pillars of leadership action for companies: (i) commitment to ethics and integrity beyond compliance; (ii) strengthening corporate culture and incentives for continuous learning and improvement; (iii) leveraging technologies;[248] (iv) supporting collective action to

---

[239] David Hess, *Partnering Against Corruption Initiative and the Business Principles for Countering Bribery*, in HANDBOOK OF TRANSNATIONAL GOVERNANCE INSTITUTION & INNOVATIONS 322 (Thomas Hale & David Held ed., 2011).

[240] David Hess, *Partnering Against Corruption Initiative and the Business Principles for Countering Bribery*, in HANDBOOK OF TRANSNATIONAL GOVERNANCE INSTITUTION & INNOVATIONS 322 (Thomas Hale & David Held ed., 2011).

[241] David Hess, *Partnering Against Corruption Initiative and the Business Principles for Countering Bribery*, in HANDBOOK OF TRANSNATIONAL GOVERNANCE INSTITUTION & INNOVATIONS 322 (Thomas Hale & David Held ed., 2011).

[242] Analyzed together, the PACI documents reveal that the first edition of the PACI Principles for Countering Corruption is from 2014; however, I did not find the document published in that year. The article uncovered a document from 2016 that, given the context, it was understood as an edition of the PACI Principles for Countering Corruption from 2014, without modifications regarding the Principles. *See*, WEF. (2016). *Partnering Against Corruption Initiative Global – Principles for Countering Corruption: Application & General Terms of Partnership.* https://www3.weforum.org/docs/WEF_PACI_Global_Principles_for_Countering_Corruption.pdf.

[243] WEF. (2013). *19th PACI Task Force Meeting Summary.* https://baselgovernance.org/sites/default/files/2019-02/wef_paci_summary_19thtaskforcemeeting.pdf.

[244] WEF. (2014). *World Economic Forum Calls on Business Leaders to Strive for Corruption-Free World.* https://www.weforum.org/press/2014/01/world-economic-forum-calls-on-business-leaders-to-strive-for-corruption-free-world/, on 30 Nov. 2023.

[245] WEF. (2016). *Partnering Against Corruption Initiative Global – Principles for Countering Corruption: Application & General Terms of Partnership.* https://www3.weforum.org/docs/WEF_PACI_Global_Principles_for_Countering_Corruption.pdf.

[246] On 30 Nov. 2023, information regarding the "Agenda for Business Integrity" was accessible on the current PACI principal page, unlike the ones about the "PACI Principles."

[247] WEF. *Overview: Partnering Against Corruption Initiative (PACI)*, https://www3.weforum.org/docs/WEF_PACI_Community_Overview_pager.pdf.

[248] The documents regarding the third pillar do not address compliance programs in specific. Regarding it, the WEF released "Hacking Corruption in the Digital Era: How Tech is Shaping the Future of Integrity in Times of Crisis" (WEF. (2020). *Hacking Corruption in The Digital Era: How Tech is Shaping the Future of Integrity in Times of Crisis: Agenda for Business Integrity.* https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_Agenda_for_Business_Integr

increase scale and impact.[249] This marks a new phase for PACI, with more extensive goals beyond promote a robust compliance program as a strategy against corruption.

i. Ethics and Integrity Beyond Compliance: Agenda for Business Integrity (2020)

The primary pillar currently guiding the PACI strategy is a conceptual foundation that calls for businesses to commit to ethics and integrity beyond compliance.[250] In essence, this entails a shift from merely preventing or reducing unethical conduct (compliance-based), typically focused on voluntary principles and standards aligned with OECD and UN Conventions, to encouraging individuals and organizations to manage business with integrity (values-based).[251] PACI emphasizes that values-based programs should evolve consistently, aligning with the progression of corporate social responsibility and business and human rights fields.[252] In this way, the programs should address not only corruption but also the high risk of human rights abuses, environmental harm, and weak rule of law in a specific market or sector, within a collective action perspective.[253]

ii. Good Intentions, Bad Outcomes? How Organizations Can Make the Leap from Box-Ticking Compliance to Building a Culture of Integrity (2020)

Concerning the second pillar of the Agenda for Business Integrity, which is to strengthen corporate culture and incentives to drive continuous learning and improvement, WEF published this document.[254] It affirms that regulators often commend robust anticorruption measures within companies. However, this dominant

---

ity_pillar_3_2020.pdf). This publication underscores technology's role in disrupting corruption, emphasizing the use of artificial intelligence to detect corrupt practices, which increases the costs of corruption for businesses while enhancing the benefits of integrity in terms of reputation and investment risk. Notably, this document addresses both businesses and governments, advocating for collaboration to leverage technology and ensure transparency, especially during crises like the COVID-19 pandemic. One output of this pillar is the Tech for Integrity, a platform that aims to accelerate anti-corruption efforts and reduce the amount of time needed to make tangible impact (WEF. *Tech for Integrity.* https://widgets.weforum.org/tech4integrity/index.html, on 30 Nov. 2023).

[249] WEF. *An Agenda for Business Integrity Four: Key Pillars of Leadership Action by Companies*. https://www3.weforum.org/docs/WEF_Pillars_English.pdf. .

[250] WEF. (2020). *Ethics and Integrity Beyond Compliance: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar1_beyond_compliance_2020.pdf.

[251] WEF. (2020). *Ethics and Integrity Beyond Compliance Global Future Council on Transparency and Anti-Corruption: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar1_beyond_compliance_2020.pdf.

[252] WEF. (2020). *Ethics and Integrity Beyond Compliance Global Future Council on Transparency and Anti-Corruption: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar1_beyond_compliance_2020.pdf.

[253] WEF. (2020). *Ethics and Integrity Beyond Compliance Global Future Council on Transparency and Anti-Corruption: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar1_beyond_compliance_2020.pdf.

[254] WEF. (2020). *Good Intentions, Bad Outcomes? How Organizations Can Make the Leap from Box-Ticking Compliance to Building a Culture of Integrity: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar2_good_intentions_bad_outcomes.pdf.

approach has led the companies' compliance team to be perceived as an internalized law enforcement body that responds to external pressure from government regulators and the public, focusing on sanctions. Nonetheless, it contends that merely penalizing individuals is insufficient, given that many employees have good intentions and grapple with navigating grey ethical lines. The document aims to elucidate how companies can motivate their employees to apply ethical reasoning in the intricate dilemmas confronting businesses, guiding them to act appropriately within the context of cultivating a culture of integrity.

The document also offers recommendations for the companies' board embrace an integrated approach by exploring the convergence of risk management, aligning ethics initiatives with the company's strategy, and prioritizing the mitigation of human rights and corruption risks. Additionally, it suggests a review of the companies' mission, strategy, and purpose in line with ESG principles, fostering ethical leadership to enhance the tone at the top, and promoting diversity and inclusion. The document also recommends measuring stakeholder trust, incorporating third-party due diligence checks – considered essential in any robust compliance program – from a complex and expensive process to a comprehensive one that includes assessments of stakeholder trust levels.

iii. Agenda for Business Integrity: Collective Action (2020)

Regarding the fourth pillar, this document emphasizes the crucial role of collaborative efforts across private, civil, and public sectors in combating corruption.[255] WEF sees collective action as a safety net for individual actors, particularly in countries with an uncertain rule of law, leveling the business playing field by uniting organizations, including the most vulnerable like SMEs, in a commitment to integrity principles.[256] It asserts that these initiatives can address governance gaps, complement legal frameworks, and prioritize practical, impactful actions over paper-based endeavors.[257]

c. The Future of Trust and Integrity (2018)

This publication is part of The Future of Trust and Integrity Project, launched in response to the 2018 financial crisis.[258] It aimed to restore trust, integrity, and address corruption by incorporating these values into systems. The Project was drawn on the Latin American region, more specifically Argentina, coinciding with the

[255] WEF. (2020). *Agenda for Business Integrity: Collective Action – Community Paper.* https://www3.weforum.org/docs/WEF_Agenda_for_Business_Integrity.pdf.

[256] The document listed four types of collective action: (i) Anti-corruption Declarations: voluntary, ethical commitments made collectively by companies, often in collaboration with civil society or the public sector; (ii) Standard-Setting Initiatives: development of sector-specific anti-corruption frameworks or standards, such as codes of ethics, to standardize integrity policies; (iv) Capacity-Building Initiatives: companies sharing resources and expertise from their compliance programs to provide training for other organizations, especially SMEs, public officials, and civil society practitioners; (v) Integrity Pacts: higher-level commitments, commonly used in public tenders, with external monitoring and certification to prevent bribery and conflicts of interest, including sanctions for non-compliance. (WEF. (2020). *Agenda for Business Integrity: Collective Action – Community Paper.* https://www3.weforum.org/docs/WEF_Agenda_for_Business_Integrity.pdf).

[257] WEF. (2020). *Agenda for Business Integrity: Collective Action – Community Paper.* https://www3.weforum.org/docs/WEF_Agenda_for_Business_Integrity.pdf.

[258] WEF. *The Future of Trust and Integrity.* https://www.weforum.org/publications/the-future-of-trust-and-integrity/, on 30 Nov. 2023.

Argentinian G20 presidency. It identified three key dimensions – institutional, behavioral, and technological – to bring positive change to corrupt systems. The institutional dimension emphasizes institution-building, rule of law enforcement, and robust compliance systems.[259]

The Future of Trust and Integrity document presents case studies demonstrating where business, government and civil society have successfully improved levels of trust and integrity to address corruption.[260] One of the cases is Integrating Comprehensive Compliance Programmes to Mitigate Corruption, highlights the success of integrating comprehensive compliance programs into a global retail company's operation, particularly in navigating diverse regulatory frameworks.[261]

In the following section, I will analyze the legal instruments mapped in this study, aiming to provide an overview of the development of direct incentives to compliance programs within the IACR.

III. AN OVERVIEW OF THE IACR: 20 YEARS OF DIRECTING PUSH FOR COMPLIANCE PROGRAMS

A. The Story Told: the IACR's Role in Corporate Liability and its Impact on Compliance Programs Spread

In the anti-corruption field, corporate liability has a landmark in the FCPA of 1977, where the United States criminalized foreign bribery.[262] From then on, the United States pressed for the expansion of the criminalization of bribery to level the playing field of regulatory standards worldwide, so that U.S. companies, already subject to the

---

[259] The behavioral dimension focuses on organizational culture, highlighting leadership, values, and effective training. The technological dimension recognizes the importance of emerging technologies like e-government, open data, and big data analytics and reflects on their implementation. The project unfolds in three phases, starting with a focus on Latin America and Africa, specifically Mexico, Argentina, and South Africa in Phase One. Phase Two extends to South Africa to establish a global framework, while Phase Three explores the Middle East or South-East Asia, accounting for sector and regional nuances. As of now, the PACI page provides information about Phase One. (WEF. *Partnering Against Corruption Initiative.* https://www.weforum.org/communities/partnering-against-corruption-initiative/, on 30 Nov. 2023).

[260] WEF. (2018). *The Future of Trust and Integrity. https://www3.weforum.org/docs/WEF_47529_The_Future_of_Trust_and_Integrity_report_2018.pdf.*

[261] PACI has other projects not directly connected to compliance programs. In 2014, PACI introduced the PACI Vanguard (WEF. (2014). *World Economic Forum Calls on Business Leaders to Strive for Corruption-Free World.* https://www.weforum.org/press/2014/01/world-economic-forum-calls-on-business-leaders-to-strive-for-corruption-free-world/, on 30 Nov. 2023). Chief executives who want to fully commitment to a higher level of leadership in anti-corruption through building trust and integrity are invited to join the PACI Vanguard (WEF. *Partnering Against Corruption Initiative.* https://www.weforum.org/communities/partnering-against-corruption-initiative/, on 30 Nov. 2023). The Vanguard's purpose is to identify innovative approaches to anti-corruption, emphasizing more meaningful dialogue and impactful collective action. Moreover, in 2020, PACI promoted the "Role and Responsibilities of Gatekeepers in the Fight Against Illicit Financial Flows: A Unifying Framework," which is a self-regulatory framework for private sector intermediaries who are strategically positioned to prevent or interrupt illicit financial flows. (WEF. (2021). *The Role and Responsibilities of Gatekeepers in the Fight Against Illicit Financial Flows: A Unifying Framework.* https://www3.weforum.org/docs/WEF_Gatekeepers_A%20Unifying_Framework_One%20pager_2021.pdf).

[262] Jennifer Arlen, *The Potentially Perverse Effects of Corporate Criminal Liability*, 23 THE JOURNAL OF LEGAL 833 (1994).

FCPA, would not compete under unequal conditions in the globalized environment.[263] This effort led to the OECD Convention, known as the "global FCPA,"[264] which came into force in 1999, mandating reforms in countries' domestic laws to criminalize bribery of foreign public officials by individuals and entities.[265] Later, the UN Convention, which entered into force in 2005, required signatory countries to establish liability for legal entities not only for bribery but also for involvement in various corruption offenses outlined in this Convention.

Other multilateral agreements prompted signatory countries to reform their legal systems to make them more rigorous against corruption.[266] The oldest one within the IACR is the OAS Convention, a regional treaty that enter into force in 1997. These conventions were a central driver that led to an impressive history of legal harmonization to make corruption unlawful and hold companies accountable for such actions.[267] The literature often suggests that the IACR, and the domestic regulations resulting from countries' adherence to these treaties, have been a strong motivation for companies to establish compliance programs, as they strengthen corporate liability and companies may adopt internal controls to reduce the risk of sanctions.[268] However, there is more to the story.

B. The Promotion of Compliance Programs Within IACR

The IACR treaties do not explicitly mention compliance programs, which are currently one of the most widespread anti-corruption strategies around the world.[269] So, how IACR play a role in the diffusion of anti-corruption compliance programs? While not discounting other possible responses, this research sheds light on the fact that the IACR has directly promoted compliance programs in non-binding documents since 2002.

Notes that how this paper was conducted,[270] no documents specifically addressing compliance programs within the African, European, EITI, FATF, Freedom House, and WTO GTA frameworks were found.[271] Furthermore, no direct incentives for compliance programs at the IMF were identified. Regarding the European Union framework, a change is on the horizon. In May 2023, a new Directive on Combating Corruption from the European Parliament and the Council was proposed and its stipulating that compliance programs will be considered a mitigating circumstance in

---

[263] *See, e.g.*, KEVIN E. DAVIS, BETWEEN IMPUNITY AND IMPERIALISM: THE REGULATION OF TRANSNATIONAL BRIBERY (2019).
[264] Jose-Miguel Bello y Villarino, *International Anticorruption Law, Revisited*, 63 HARVARD INTERNATIONAL LAW JOURNAL 343 (2022), at 351.
[265] OECD (1997). *Convention on Combating Bribery of Foreign Public Officials in International Business Transactions*. https://www.oecd.org/daf/anti-bribery/ConvCombatBribery_ENG.pdf.
[266] SUSAN ROSE-ACKERMAN & BONNIE. J. PALIFKA, CORRUPTION AND GOVERNMENT: CAUSES, CONSEQUENCES, AND REFORM (2016).
[267] Erling Hjelmeng & Tina Søreide, *Bribes, Crimes and Law Enforcement*, 28 EUROPEAN BUSINESS LAW REVIEW, 19 (2017).
[268] *See, e.g.,* Kevin E. Davis & Veronica R. Martinez, *Transnational Anti-bribery Law*, in THE CAMBRIDGE HANDBOOK OF COMPLIANCE 924 (Benjamin van Rooij & D. Daniel Sokol ed., 2021).
[269] OECD. (2020). *Corporate Anti-Corruption Compliance Drivers, Mechanisms, and Ideas for Change*. https://www.oecd.org/corruption/corporate-anti-corruption-compliance.htm, on 20 Jan. 2024.
[270] *See* Section 2, footnotes 34, 37, 38, 39, 40, 41, 42, 43, 44, and 45.
[271] These actors were included in the IACR by one or more authors in the literature analyzed, *see* Section 2.

cases of the offense under the terms of the Directive.[272] From now on, I will analyze the several direct incentives to compliance programs found within the IACR.

1. The Targets

The 52 documents that directly promote compliance within the IACR, mapped in this study, have varied targets: most of them focus on companies (63%), some on governments (19%), some on both governments and companies (10%), and others on collective actions (8%). The documents targeting companies generally emphasize the idea that companies should oppose corruption not just because it is illegal but also because it is beneficial for businesses.[273] Among the documents examined, the earliest one promoting corporate adoption of compliance program is from 2002 and was authored by TI. This document outlines the minimum requirements that a company should follow in implementing an effective anti-bribery compliance program. All analyzed actor categories have documents encouraging companies to adopt compliance programs. Most of these documents were produced by private international initiatives (50%).

In addition to those targeting companies, I found within the IACR documents that emphasize compliance programs as a tool in collective actions (8% of the total analyzed).[274] They often link collective action to SMEs, acknowledging that these companies cannot combat corruption alone in a market where larger companies act corruptly. Collective actions can involve both private and public actors aiming to establish anti-corruption standards that companies commit to actively comply with, leveling the playing field for a group of companies or a sector of the market. The oldest

---

[272] Article 18 stipulates that mitigating circumstances will be considered "where the offender is a legal person and it has implemented effective internal controls, ethics awareness, and compliance programs to prevent corruption prior to or after the commission of the offense." (EUR-Lex. *Document 52023PC0234*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2023%3A234%3AFIN, on 6 Dez. 2023).

[273] For instance, UN Global Compact, TI, International Business Leaders Forum. (2005). *Business Against Corruption: A Framework for Action – Implementation of the 10th UN Global Compact Principle Against Corruption*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2F7.7%2FBACtextcoversmallFINAL.pdf; ICC, TI, UN Global Compact, and PACI. (2008). *Clean Business is Good Business*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2Fclean_business_is_good_business.pdf; UN. (2009). *Global Compact for the 10th Principle: Corporate Sustainability with Integrity*. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FUNGC_AntiCorruptionReporting.pdf;

[274] UNIDO & UNODC. (2007). *Corruption Prevention to Foster Small and Medium-Sized Enterprise Development: Providing Anti-Corruption Assistance to Small Businesses in The Developing World.* https://www.unodc.org/documents/corruption/Publications/2012/UNIDO-UNODC_Publication_on_Small_Business_Development_and_Corruption_Vol1.pdf; UN Global Compact. (2012). *Global Compact for the 10th Principle: Corporate Sustainability with Integrity – Organizational Change to Collective Action.* https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FGC_for_the_10th_Principle.pdf; UNIDO & UNODC. (2012). *Corruption Prevention to Foster Small and Medium-Sized Enterprise Development.* https://www.unodc.org/documents/corruption/Publications/2012/Corruption_prevention_to_foster_small_and_medium_size_enterprise_development_Vol_2.pdf; WEF. (2020). Agenda for Business Integrity: Collective Action – Community Paper. https://www3.weforum.org/docs/WEF_Agenda_for_Business_Integrity.pdf.

document that I located focusing on collective action was published by the UN in 2007.[275]

The mapping also reveals documents recommending that governments reform their legal system to adopt legal incentives to encourage companies to implement compliance programs. Among those, 60% were produced by international organizations, 34% by intergovernmental initiatives, and 6% by private international initiatives. The oldest one was published by ICC in 2005, recommending that states established compliance programs as a requirement for government large contracts.[276] In 2009, the OECD recommended that governments grants benefit in public binding process,[277] and, in 2015, to require compliance in some governments contracts.[278] OECD also published another document that suggest governments to stimulate compliance programs without specify some strategy in 2016.[279] In 2021, a OECD document recommended the following strategies among that one located in the countries studied: reduce penalty, grants benefit in public binding process, and impacts the decision to prosecute. Regarding UN, in 2013, the UNODC published two documents. The first one recommended that governments adopt a strategy to promote compliance programs: eliminate liability; clause in non-trial resolution whit the state, requirement for governments contractors; reduce penalty; grants benefit in public binding process and as a requirement for mitigating or lifting debarment.[280] The second one, specific to Indian context, suggesting compliance programs mandatory for certain business and as a penalty.[281] Another UN document stimulate governments to stimulate corporate compliance programs in a connection whit human rights.[282] Whitin

---

[275] UNIDO & UNODC. (2007). *Corruption Prevention to Foster Small and Medium-Sized Enterprise Development: Providing Anti-Corruption Assistance to Small Businesses in The Developing World.* https://www.unodc.org/documents/corruption/Publications/2012/UNIDO-UNODC_Publication_on_Small_Business_Development_and_Corruption_Vol1.pdf.

[276] ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf.

[277] OECD. (2009). *OECD Recommendation for Further Combating Bribery of Foreign Public Officials in International Business Transactions*. https://www.oecd.org/investment/anti-bribery/anti-briberyconvention/oecdantibriberyrecommendation2009.htm#:~:text=About%20the%202009%20Recommendation&text=The%20Recommendation%20was%20adopted%20by,Internal%20Controls%2C%20Ethics%20and%20Compliance.

[278] OECD. (2015). *Recommendation of the Council on Public Procurement.* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0411.

[279] OECD. (2016). *Ministerial Declaration – The Fight Against Foreign Bribery: Towards a New Era of Enforcement*. https://www.oecd.org/corruption/OECD-Anti-Bribery-Ministerial-Declaration-2016.pdf.

[280] UNODC. (2013). *A Resource Guide on State Measures for Strengthening Corporate Integrity*. https://www.unodc.org/documents/corruption/Publications/2013/Resource_Guide_on_State_Measures_for_Strengthening_Corporate_Integrity.pdf.

[281] UNODC. (2013). *Corporate Integrity: Incentives for Corporate Integrity in Accordance with the United Nations Convention Against Corruption – A Report*. https://www.unodc.org/documents/southasia//publications/research-studies/CI_Report.pdf

[282] UN. (2020). *Connecting the Business and Human Rights and the Anti-Corruption Agendas: Report of the Working Group on the Issue of Human Rights and Transnational Corporations and other Business Enterprises (A/HRC/44/43)*. https://digitallibrary.un.org/record/3889182, on 6 Dec. 2023.

G20, generic documents also were published in 2014,[283] 2015,[284] 2019,[285] and 2021.[286] A document dated from 2017 specifically suggest that governments adopt the strategies of grants benefit in public binding process and reduce penalty.[287] Finally, within OAS, two states members initiative to promote compliance programs as a public recognition of companies that implement it are recognized as a good practice.[288] Table 1 below summarizes the strategies recommended by the IACR for governments to adopt in promoting corporate compliance programs.

[283] UNODC. (2014). *G20 Anti-corruption Action Plan 2015-2016*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2014_G20_ACWG_Action_Plan_2015-2016.pdf.

[284] UNODC. (2015). *G20 Principles for Promoting Integrity in Public Procurement*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Public-Sector-Integrity-and-Transparency/G20-Principles_for_Promoting_Integrity_in_Public_Procurement_2015.pdf.

[285] UNODC. (2019). *G20 Compendium of Good Practices for Promoting Integrity and Transparency in Infrastructure Development.* https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Sectors/G20_Compendium_of_Good_Practices_for_Promoting_Integrity_and_Transparency_in_Infrastructure_Development_2019.pdf.

[286] UNODC. (2021). *G20 Anti-corruption Action Plan 2022-2024*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2021_G20_Anti-Corruption_Action_Plan_2022-2024.pdf.

[287] UNODC. (2017). *G20 High-Level Principles on the Liability of Legal Persons for Corruption*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Bribery/G20_High_Level_Principles_on_the_Liability_of_Legal_Persons_for_Corruption_2017.pdf.

[288] OAS. *Anticorruption Portal of the Americas – Best Practices to Prevent and Combat Corruption.* http://www.oas.org/en/sla/dlc/mesicic/buenas-practicas.html, on 19 Jul. 2023.

Table 1: Legal strategies recommended by IACR actors for governments to promote compliance programs

Legal strategy	**2005**	**2009**	**2013**	**2014**	**2015**	**2016**	**2017**	**2019**	**2020**	**2021**	**2023**
not specified				G20	G20	OECD		G20	G20	G20	
eliminates liability			UN								
reduce penalty			UN				G20			OECD	
clause in non-trial resolutions with the state			UN								
mandatory for certain businesses			UN								
requirement for government contractors	ICC		UN		OECD						
penalty			UN								
grants benefit in the public bidding process		OECD	UN				G20			OECD	
requirement for mitigating or lifting debarment			UN								
impacts the decision to prosecute										OECD	
public recognition								OAS			OAS

Note that both the first located documents encouraging companies (2002) and those targeting governments (2005) to stimulate compliance programs were produced by private international initiatives (TI and ICC, respectively). Private initiatives are also responsible for a significant portion of the located documents promoting compliance programs (35%). This indicates that the private sector plays a relevant role in this field and also suggests an interest from companies in the diffusion of compliance programs. This aligns with what happened in the United States previously, concerning the legal incentive for compliance set forth by the United States Sentencing Commission (USSC) in 1991. Nolan Ezra Clark asserts that this incentive, entailing a reduction in sanctions for organizations with compliance programs, was driven by a lobby of companies interested in a strategy to decrease possible sanctions in the face of stricter rules against corruption.[289]

2. Incentives for Governments to Promote Compliance Programs: A Comparison of IACR and Some Domestic Legal Reforms

The origin of compliance is contested, however, there is relative consensus that the first domestic reform aimed at creating direct legal incentives for legal persons to adopt compliance programs took place in the United States in 1991.[290] The USSC, the institution responsible for determining the sentencing guidelines for U.S. federal courts by means of the Federal Sentencing Guidelines Manual, first provided for the reduction of penalties applied to organizations that established patterns and procedures of compliance at Sentencing Guidelines for Organizations (SGO), being this the paradigmatic incentive to compliance programs.[291] The SGO also allowed courts to impose, in specific cases, a compliance program to prevent and detect violations of law as a condition of probation.[292] The literature identifies the SGO as a milestone in the "era of compliance," where compliance programs play a strategic role in companies, with a specialized industry actively supporting this strategy.[293]

Following the OSG, federal prosecutors started providing corporate defendants with settlements that considered compliance programs.[294] To standardize this approach, in 1999, the Department of Justice (DoJ) Deputy Attorney General Eric H. Holder issued a policy memorandum entitled "Federal Prosecution of Corporations" (the

---

[289] Nolan Ezra Clark, *Corporate Sentencing Guidelines: Drafting History* (updated by the Editors), in COMPLIANCE PROGRAMS AND THE CORPORATE SENTENCING GUIDELINES: PREVENTING CRIMINAL AND CIVIL LIABILITY (William M. Hannay ed., 2022).

[290] *E.g.*, Sean. J. Griffith, *Corporate Governance in an Era of Compliance*, 57 WILLIAM AND MARY LAW REVIEW 2075 (2016); Eugene F. Soltes, *Evaluating the Effectiveness of Corporate Compliance Programs: Establishing a Model for Prosecutors, Courts, and Firms*, 14 NEW YORK UNIVERSITY JOURNAL OF LAW & BUSINESS 965 (2018).

[291] §8C2.5.(f) (USSC. (1991). *1991 Federal Sentencing Guidelines Manual*. (1991, Nov.). USSC, Washington. Retrieved June 30, 2022, from https://www.ussc.gov/guidelines/archive/1991-federal-sentencing-guidelines-manual).

[292] §8D1.4.(c)(1) (USSC. (1991). *1991 Federal Sentencing Guidelines Manual*. (1991, Nov.). USSC, Washington. Retrieved June 30, 2022, from https://www.ussc.gov/guidelines/archive/1991-federal-sentencing-guidelines-manual).

[293] David Hess, *Ethical Infrastructure and Evidence-Based Corporate Compliance and Ethics Programs: Policy Implications from the Empirical Evidence*, 12 NEW YORK UNIVERSITY JOURNAL OF LAW AND BUSINESS 317 (2016).

[294] Sean. J. Griffith, *Corporate Governance in an Era of Compliance*, 57 WILLIAM AND MARY LAW REVIEW 2075 (2016); Eugene F. Soltes, *Evaluating the Effectiveness of Corporate Compliance Programs: Establishing a Model for Prosecutors, Courts, and Firms*, 14 NEW YORK UNIVERSITY JOURNAL OF LAW & BUSINESS 965 (2018).

"Holder Memorandum").[295] This document recommends prosecutors to consider in conducting an investigation, determining whether to bring charges, negotiating plea agreements, and reaching a decision as to the proper treatment of a corporate target, the corporation's remedial actions, including any efforts to implement an effective corporate compliance program or to improve an existing one.[296] This a stance reinforced in 2019 by the DoJ "Evaluation of Corporate Compliance Programs."[297]

The United States continued to directly encourage compliance programs through subsequent legal developments. Until 2007, ethics programs and practices of defense contractors were self-policed.[298] Given the significant sums of federal dollars spent by agencies to acquire goods and services and the need to establish a clear and consistent policy regarding the contractor code of ethics and business conduct, the Federal Acquisition Regulation (FAR) was amended in 2007 to mandate contractor ethics programs.[299] The 2021 version, currently in force, establishes as a general rule that contractors must implement an ongoing business ethics awareness and compliance program within 90 days of contract award.[300]

Thus, the United States adopted the strategy of promoting compliance programs by reducing penalties and imposing compliance programs as a condition of probation in 1991. Within the IACR, these strategies only appeared in 2013. Moreover, in 2009, the United States formalized that compliance programs should be taken into account by prosecutors in their decisions regarding prosecution or non-trial resolution, strategies first recommended within the IACR later, in 2021 and 2013, respectively. In contrast, the United States introduced the strategy of requiring compliance programs in some government contracts in 2007, whereas the IACR made this recommendation in 2005.

This analysis can also be extended to other countries. For instance, in 2001, Italy enacted the Legislative Decree 231 which stipulated that a compliance program adopted before an unlawful act can eliminate corporate liability if it meets regulatory

---

[295] Sean. J. Griffith, *Corporate Governance in an Era of Compliance*, 57 WILLIAM AND MARY LAW REVIEW 2075 (2016); Eugene F. Soltes, *Evaluating the Effectiveness of Corporate Compliance Programs: Establishing a Model for Prosecutors, Courts, and Firms*, 14 NEW YORK UNIVERSITY JOURNAL OF LAW & BUSINESS 965 (2018); Joshi Attorneys and Counselors. *An End to "Backseat Driving"? The Thompson Memorandum and Government Tactics in White-Collar Crime Investigation and Prosecution.* https://www.joshiattorneys.com/articles-and-publications/an-end-to-backseat-driving-the-thompson-memorandum-and-government-tactics-in-white-collar-crime-investigation-and-prosecution/, on 22 Jan. 2024.

[296] DoJ. (1999). *Bringing Criminal Charges Against Corporations.* https://www.justice.gov/sites/default/files/criminal-fraud/legacy/2010/04/11/charging-corps.PDF.

[297] O'Shea, A. et al. DOJ Updates Guidance on the Evaluation of Corporate Compliance Programs. *Harvard Law School Forum on Corporate Governance*, 20 Jun. 2020. https://corpgov.law.harvard.edu/2020/06/20/doj-updates-guidance-on-the-evaluation-of-corporate-compliance-programs/#12.

[298] GAO. (2009). *Report to Congressional Committees: Defense Contracting Integrity – Opportunities Exist to Improve DoD's Oversight of Contractor Ethics Programs*. https://www.gao.gov/assets/gao-09-591.pdf.

[299] U.S. Federal Register. (2007). *Contractor Code of Business Ethics and Conduct*. https://www.govinfo.gov/content/pkg/FR-2007-11-23/pdf/07-5800.pdf.

[300] Code of Federal Regulation. *Title 48, Chapter 1, Subchapter H, Part 52, Subpart 52.2, 52.203-13: Contractor Code of Business Ethics and Conduct (Nov. 2021)*, on 23 Jan. 2024. https://www.ecfr.gov/current/title-48/chapter-1/subchapter-H/part-52/subpart-52.2/section-52.203-13.

requirements[301] – a recommendation found within the IACR only in 2013. In contrast, Brazil reform its laws to grants benefits in the public bidding process for companies with compliance programs in 2021,[302] after the IACR made this recommendation in 2005, for the first time. Moreover, both Colombia[303] and France,[304] reformed their legal system in 2016 to make compliance programs mandatory for certain business, while IACR did this recommendation early, in 2013.

This research sheds light on the possibility of new studies analyzing the relationship between domestic statutes and the IACR regarding legal incentives for compliance programs. What can be asserted at this point is that legal incentives for compliance programs by states emerged earlier in local frameworks (with the paradigmatic legal development occurring in the United States in 1991) than in the IACR (which, according to this research, first encouraged governments to adopt strategies to promote corporate compliance programs in 2005).

3. The Compliance Industry

For over a decade, scholars have debated how the U.S. adjudicative model of compliance regulation has not only benefited the compliance programs industry but has also played a role in its creation.[305] Presently, the compliance industry boasts a robust presence. It has its own professional organizations and constitutes a distinct legal practice area,[306] with universities offering courses and degrees tailored specifically to compliance.[307] In parallel, compliance departments within many organizations have expanded both in size and significance.[308]

This article reveals that the IACR stimulates the compliance industry. This mapped found explicit suggestions for companies to hire third parties to implement, monitor, and/or certify their compliance programs within the IACR documents. For instance, the 2009 Business Principles for Countering Bribery led by TI suggests that companies should consider commissioning external verification or assurance to

---

[301] Article 6. (Italy. (2001). *Decreto Legislativo 8 giugno 2001, n. 231.* https://www.normattiva.it/atto/caricaDettaglioAtto?atto.dataPubblicazioneGazzetta=2001-06-19&atto.codiceRedazionale=001G0293&atto.articolo.numero=0&atto.articolo.sottoArticolo=1&atto.articolo.sottoArticolo1=10&qId=aaa75281-73a8-48f4-a506-9b9755533a67&tabID=0.9434481816169746&title=lbl.dettaglioAtto, on 30 Aug. 2023).

[302] Article 60, IV (Brazil. (2021). *Lei 14.133, de 1º de abril de 2021.* Retrieved June 30, 2022, from http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm).

[303] Article 23 (Colombia. (2016). *Ley 1778 de 2016*. Retrieved June 30, 2022, from https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=67542).

[304] Article 17. (France. (2016). *Loi no. 2016-1691 du 9 décembre 2016 relative à la transparence, à la lutte contre la corruption et à la modernisation de la vie économique*. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033558528).

[305] *See, e.g.,* Mirian H. Baer, *Governing Corporate Compliance*, 50 BOSTON COLLEGE LAW REVIEW 949 (2009).

[306] Geoffrey P. Miller, *Compliance: Past, Present and Future*, 48 UNIVERSITY OF TOLEDO LAW REVIEW 437 (2017).

[307] D. Daniel Sokol, *Teaching Compliance*, 84 UNIVERSITY OF CINCINNATI LAW REVIEW 399 (2016).

[308] Benjamin van Rooij & D. Daniel Sokol, *Introduction: Compliance as the Interaction between Rules and Behavior*, in THE CAMBRIDGE HANDBOOK OF COMPLIANCE 1 (Benjamin van Rooij & D. Daniel Sokol ed., 2021).

enhance internal and external confidence in the compliance program's effectiveness.[309] The Global Compact for the 10th Principle: Corporate Sustainability with Integrity, published in 2012, asserts that certification helps companies continuously improve their compliance programs based on the recommendations given by external experts.[310] Also published in 2012, the TI Assurance Framework for Corporate Anti-bribery Programs affirms that third-party review enhances credible anti-bribery systems' demand.[311] The G20 High-Level Principles on Private Sector Transparency and Integrity, from 2015, suggest that companies may seek professional advice to learn more about the compliance program most appropriate for their business.[312] It is widely acknowledged that a robust market for compliance programs has enhanced in the last decades.[313] This paper structure does not permit an inference as to whether these documents have influenced the growth of this market or vice versa.

4. From Anti-Bribery to ESG

This article mapping of the development of compliance programs within the IACR over the last two decades, especially the analysis of documents revised over time, such as the Guidelines for Multinational Enterprises[314] and the ICC Rules,[315] offers an overview of the changes in how the IACR approaches compliance programs over time. In general liner, initially, the IACR focusing on incentives to compliance programs, to fight bribery, specifically bribery in transnational business, aligned with the OECD Convention. Over time, new initiatives have emerged to suggests that the compliance programs should address a broader spectrum of corrupt activities, in line whit the UN Convention. Newer documents propose an even more extensive range of objectives for anti-corruption compliance programs. The G20 Anti-corruption Action Plan 2022-2024, for instance, suggests that the analyses and solutions against corruption should consider the impacts of corruption on cross-cutting issues such as environmental and gender-

---

309 TI. (2013). *Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International.* https://www.transparency.org/en/publications/business-principles-for-countering-bribery.

310 UN Global Compact. (2012). *Global Compact for the 10th Principle: Corporate Sustainability with Integrity –Organizational Change to Collective Action.* https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2FAnti-Corruption%2FGC_for_the_10th_Principle.pdf.

311 TI. (2012). *Assurance Framework for Corporate Anti-Bribery Programmes.* https://knowledgehub.transparency.org/product/assurance-framework-for-corporate-anti-bribery-programmes.

312 UNODC. (2015). *G20 High-Level Principles on Private Sector Transparency and Integrity.* https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Thematic-Areas/Private-Sector-Integrity-and-Transparency/G20_High_Level_Principles_on_Private_Sector_Transparency_and_Integrity_2015.pdf.

313 *See, e.g.,* Eugene F. Soltes, *Evaluating the Effectiveness of Corporate Compliance Programs: Establishing a Model for Prosecutors, Courts, and Firms*, 14 NEW YORK UNIVERSITY JOURNAL OF LAW & BUSINESS 965 (2018).

314 OECD. (2011). OECD *Guidelines for Multinational Enterprises.* https://www.oecd.org/daf/inv/mne/48004323.pdf; OECD. (2023). OECD *Guidelines for Multinational Enterprises on Responsible Business Conduct*. https://doi.org/10.1787/81f92357-en.

315 More specifically, ICC. (2005). *Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations*. https://iccwbo.org/wp-content/uploads/sites/3/2005/10/Combating-Extortion-and-Bribery-ICC-Rules-of-Conduct-and-Recommendations.pdf, and the subsequent version ICC. (2011). *ICC Rules on Combating Corruption.* https://iccwbo.org/wp-content/uploads/sites/3/2011/10/ICC-Rules-on-Combating-Corruption-2011.pdf.

related matters.[316] The 2020 UN report, Connecting the Business and Human Rights and the Anti-corruption Agendas, suggests that anti-corruption programs should have an integrated approach to responsible business conduct, including considerations for human rights.[317] In the same year, WEF published the document Ethics and Integrity Beyond Compliance, stimulating that compliance programs go beyond corruption and address ESG issues.[318]

The change regarding the approach of compliance programs from an anti-bribery to an ESG approach is evident when comparing documents from different decades. For instance, the TI Business Principles for Countering Bribery from 2014 recommends a program that balances a compliance-based and values-based approaches,[319] whereas the current PACI directive, from 2020, urges the adoption of compliance program with a value-based approach.[320] These developments seem to reflect an ongoing effort to adapt compliance programs to changing contexts and emerging challenges, given that despite the diffusion of compliance programs, their ability to achieve what they propose is still debated.[321] It also can reflect the aim of some actors to use a more well-known instrument, the compliance programs, to give more legitimacy to the ESG approach – a topic currently under intense discussion.[322]

C. The IACR Before the Compliance Programs

While a connection exists between the development of the IACR and the incentive to compliance programs within it, this article highlights that references to compliance programs within the IACR have not been prevalent since its inception. Compliance programs emerged as a strategy within the IACR sometime after its origin, marked by the OAS Convention in 1997. As revealed in this article, the first incentive for compliance programs found within the IACR is from 2002, and the explosion of these incentives in the IACR occurs in the 2010s.[323] Some documents analyzed suggest

---

316 See, *e.g.*, UNODC. (2021). *G20 Anti-corruption Action Plan 2022-2024*. https://www.unodc.org/documents/corruption/G20-Anti-Corruption-Resources/Action-Plans-and-Implementation-Plans/2021_G20_Anti-Corruption_Action_Plan_2022-2024.pdf.

317 UN. (2020). *Connecting the Business and Human Rights and the Anti-Corruption Agendas: Report of the Working Group on the Issue of Human Rights and Transnational Corporations and other Business Enterprises (A/HRC/44/43)*. https://digitallibrary.un.org/record/3889182, on 6 Dec. 2023.

318 WEF. (2020). *Ethics and Integrity Beyond Compliance: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar1_beyond_compliance_2020.pdf.

319 TI. (2004). *Business Principles for Countering Bribery: Guidance Document.* https://www.ethics.org/wp-content/uploads/resources/Business_Principles_for_Countering_Bribery_Transparency_Intl_Guidance_Document_2004.pdf.

320 WEF. (2020). *Ethics and Integrity Beyond Compliance Global Future Council on Transparency and Anti-Corruption: Agenda for Business Integrity*. https://www3.weforum.org/docs/WEF_GFC_on_Transparency_and_AC_pillar1_beyond_compliance_2020.pdf.

321 About the debate *see, e.g.*, Benjamin van Rooij & D. Daniel Sokol, *Introduction: Compliance as the Interaction between Rules and Behavior*, in THE CAMBRIDGE HANDBOOK OF COMPLIANCE 924 (Benjamin van Rooij & Sokol ed., 2021).

322 *See*, *e.g.*, Simon Watkins, The ISSB's Battle to Sort the Alphabet Soup of ESG Reporting. *Financial Times*, https://professional.ft.com/en-gb/blog/the-issbs-battle-to-sort-the-alphabet-soup-of-esg-reporting/, on 27 Dec. 2023.

323 I found 52 documents within the IACR that directly endorse compliance programs, published between 2002 and 2023. The number of publications per year was: 1 (2002); 1 (2003); 3 (2004); 2

that compliance programs were an approach to countering corruption by companies, giving concreteness to the anti-corruption convention aims.[324]

CONCLUSION

We are living in the era of corporate compliance programs. Despite the relevance of this programs to businesses, governments, and societies, there is a dearth of literature on the driving forces and mechanisms that influence their development. The anti-corruption literature commonly associates the proliferation of compliance programs with the reinforcement of global anti-corruption laws, especially through conventions that stand corporate liability for corrupt practices. However, these conventions do not explicitly mention compliance programs. So, how did the International Anti-Corruption Regime (IACR) contribute to the explosion of compliance programs around the world?

This innovative article mapped documents produced by 18 international actors – including international organizations, international financial institutions, intergovernmental initiatives, and private international initiatives – seeking insights into their treatment of corporate compliance programs. I found 52 documents that expressly promote compliance programs. These direct mentions are in some non-binding instruments published after 2002 rather than in international conventions. Most of these documents promote compliance programs by encouraging businesses to adopt them, affirming that this is a good strategy for business. There are also documents targeting governments to create legal incentives for companies to adopt these programs starting from 2005, understanding that governments should promote corporate compliance programs as a public strategy against corruption. Others emphasize collective action involving groups of companies alongside various actors, suggesting that fighting corruption is a challenge that should be faced by multiple actors, including civil society.

This article concludes that besides the IACR's role in leveling the playing field regarding corporate liability for corruption, it has also elected corporate compliance programs as a strategy against corruption, which must be promoted by governments, adopted by companies, and supported by civil society. This article also demonstrates that there is a change in the approach of the model of compliance programs stimulated by these documents in the last two decades. First, the IACR promoted compliance programs specifically focused on bribery, later broadening the scope to address corruption more generally, and most recently, suggesting an ESG approach.

This study also reveals that the IACR began incentivizing compliance programs as a strategy against corruption in 2002, a long time after this regime's initial inception, which dates back to the OAS Convention in 1996. Regarding the recommendations within the IACR for governments to reform the legal system to promote corporate compliance programs, first found in 2005 within this regime, it lags behind the domestic paradigmatic legal reform to stimulate corporate compliance, which took place in the

(2005); 1 (2007); 3 (2008); 4 (2009); 2 (2010); 4 (2011); 3 (2012); 5 (2013); 2 (2014); 5 (2015); 2 (2016); 1 (2017); 2 (2018); 2 (2019); 4 (2020); 2 (2021); 3 (2023).

[324] *See*, *e.g.*, TI. (2004). *Business Principles for Countering Bribery: Guidance Document.* https://www.ethics.org/wp-content/uploads/resources/Business_Principles_for_Countering_Bribery_Transparency_Intl_Guidance_Document_2004.pdf.

United States in 1991. The IACR's promotion of compliance programs appears to mirror an external movement of the rise of this mechanism. However, this does not mean that the IACR does not have the potential to influence states to reform the legal system to promote compliance programs. As illustrated in this article, some countries adopted incentives for compliance programs after the IACR's recommendations in this way. Future studies can delve into investigating the relationship between the IACR and domestic regulations regarding legal incentives for compliance programs.

This paper defines, maps, and analyzes the IACR, focusing on compliance programs, contributing to an understanding of the explosion of this strategy around the world, as well as to the development of the field of compliance program studies in international scholarship.

APPENDIX

Table 2: Recommendations on compliance programs within the IACR

Year	#*	Document	Actor(s)	Focus on	What does it say? (Emphasis add)
2002	1	Business Principles for Countering Bribery: An Initiative of Transparency International and Social Accountability International	TI and Social Accountability International	Companies	"The Business Principles The enterprise shall prohibit bribery in any form whether direct or indirect **The enterprise shall commit to implementation of a Programme to counter bribery**" (p. 2)
2003	2	Business Principles for Countering Bribery: An Essential Tool	TI	Companies	Reissue of the 2002 document.
2004	3	Business Principles for Countering Bribery: Guidance Document	TI	Companies	"Business Principles for Countering Bribery: Six Step Implementation 1. Decide to adopt an anti-bribery policy; 2. Plan implementation; 3. **Develop anti-bribery programme tailored to the business**; 4. Implement; 5. Monitor; 6. Evaluate." (p. 61)

	4	OECD Principles of Corporate Governance (second edition)	OECD	Companies	"**Companies are also well advised to set up internal programmes and procedures to promote compliance with applicable laws, regulations and standards, including statutes to criminalise bribery of foreign officials** that are required to be enacted by the OECD Anti-bribery Convention and measures designed to control other forms of bribery and corruption. Moreover, compliance must also relate to other laws and regulations such as those covering securities, competition and work and safety conditions. **Such compliance programmes will also underpin the company's ethical code**. To be effective, the incentive structure of the business needs to be aligned with its ethical and professional standards so that adherence to these values is rewarded and breaches of law are met with dissuasive consequences or penalties. **Compliance programmes should also extend where possible to subsidiaries**." (p. 63)
	5	PACI Principles for Countering Bribery	WEF, TI and the Basel Institute on Governance.	Companies	"**The enterprise shall commit to the continuation or implementation of an effective Program to counter Bribery**. An effective Program is the entirety of an enterprise's anti-bribery efforts, specifically including its code of ethics, policies and procedures, administrative processes, training, guidance and oversight. This commitment is to develop and administer an internal compliance Program that effectively makes an enterprise's anti-corruption policy an integral part of daily practice." (p. 6)
2005	6	Business Against Corruption: A Framework for Action - Implementation of the 10th UN Global Compact Principle Against Corruption (first edition)	UN (Global Compact), TI, International Business Leaders Forum	Companies	"The UN Global Compact suggests to participants to consider the following three elements when **fighting corruption and implementing the 10th principle**. Internal: **As a first and basic step, introduce anti-corruption policies and programmes within their organisations and their business operations**; External: Report on the work against corruption in the annual Communication on Progress; and share experiences and best practices through the submission of examples and case stories; Collective: Join forces with industry peers and with other stakeholders." (p. 8) "**Transparency International** has developed a Six-Step Implementation Process based on the Business Principles for Countering Bribery. This practical **guide assists companies in developing and implementing an anti-bribery policy**. The TI Six-Step Implementation Process can be modified to take into account the size of a company and its ability to complete the steps

					within the suggested timeframe." (p. 11)
	7	Combating Extortion and Bribery: ICC Rules of Conduct and Recommendations (fourth edition of ICC Rules)	ICC	Companies and Governments	"ICC has emphasized the critical role of compliance by enterprises with self-imposed rules based on their own values, while recognizing the basic responsibility of national governments and international organizations in the fight against corruption. Adhering to strict rules defined within the enterprise will help businesses fulfill their legal obligations in a more natural and effective way. **The adoption and implementation of their own integrity programs is, therefore, strongly recommended**." (p. 2) "**Adoption of antibribery compliance programmes should be a condition for bidding on major government contracts**." (p. 12)
2007	8	Corruption Prevention to Foster Small and Medium-Sized Enterprise Development: Providing Anti-corruption Assistance to Small Businesses in the Developing World (first edition)	UN (UNIDO and UNODC)	Collective action	"Even though internal measures are usually implemented more easily and quicker in SMEs than in large companies, **internal codes of conduct and compliance programmes alone are in many cases not helpful for SMEs**, as they usually lack either the resources or the market power to stand by their zero-tolerance policies. In particular, they risk being driven out of their market by competitors that do not adhere to such standards. One way to support those companies that do not have the **power to tackle the problem alone is collective action**." (p. 17)
2008	9	Business Principles for Countering Bribery: Small and Medium Enterprise (SME) Edition	TI	Companies	"This version is a simplification of the processes written to help smaller organisations with fewer resources, through clarification of the issues and practical examples. The values held by the Business Principles are unchanged." (p. 7) "**These guidelines are intended to help you implement your anti-bribery Programme which addresses your structure, business and risks**." (p. 28)

	10	Clean Business is Good Business: The Business Case against Corruption	ICC, TI, UN (Global Compact), PACI	Companies	"What Can Your Company Do? **An increasing number of companies are demonstrating leadership by implementing effective anti-corruption programmes within their companies**. Common features of such programmes include: […]" (p. 2)
	11	Fighting Corruption: International Corporate Integrity Handbook	ICC	Companies	"Setting up a compliance programme will help an enterprise make its code a reality, and it will also be in its self-interest. For example, in a number of countries, a fully fledged compliance plan can help the enterprise show that it has used all reasonable means to avoid prohibited behavior." (p. 79)
2009	12	Business Principles for Countering Bribery: A Multi-Stakeholder Initiative led by Transparency International	TI	Companies	"Enterprises should implement anti-bribery programmes both as an expression of core values of integrity and responsibility, but also to counter the risk of bribery. Risk will vary across different industries and specific companies, but no enterprise can be sure that that it will be free of risk. **An effective anti-bribery programme strengthens reputation, builds the respect of employees, raises credibility with key stakeholders and supports an enterprise's commitment to corporate responsibility**." (p. 5)
	13	Business Principles for Countering Bribery: Transparency International Self-Evaluation Tool	TI	Companies	"**This Self-Evaluation Tool (SET) has been developed by Transparency International for use by companies to self-evaluate their anti-bribery Programmes**. It aims to enable companies to appraise the strength, completeness and effectiveness of their anti-bribery policies and procedures against the framework of the Business Principles for Countering Bribery." (p. 3)
	14	Reporting Guidance on the 10th Principle Against Corruption	UN (Global Compact) and TI	Companies	"The adoption of the 10th Principle sent a strong worldwide signal that the private sector and other non-state-actors share responsibility for eliminating corruption and stand ready to play their part. **The 10th Principle commits Global Compact participants not only to avoid bribery, extortion and other forms of corruption, but also to develop policies and concrete programmes to address it**. Companies are challenged to join governments, UN agencies and civil society to realize a more transparent global economy." (p. 5)
	15	Recommendation of the Council for	OECD	Companies and Governme	"**Member countries should encourage**: i) **Companies to develop** and adopt adequate

		Further Combating Bribery of Foreign Public Officials in International Business Transactions (first edition)		nts	internal controls, ethics and **compliance programmes** or measures for the purpose of preventing and detecting foreign bribery, taking into account the Good Practice Guidance on Internal Controls, Ethics, and Compliance, set forth in Annex II hereto, which is an integral part of this Recommendation; […] vi) **Their government agencies to consider, where international business transactions are concerned, and as appropriate, internal controls, ethics, and compliance programmes or measures in their decisions to grant public advantages,** including public subsidies, licences, public procurement contracts, contracts funded by official development assistance, and officially supported export credits." (p. 5) "Annex II: Good practice guidance on internal controls, ethics, and compliance […] **This Good Practice Guidance (hereinafter "Guidance") is addressed to companies for establishing and ensuring the effectiveness of internal controls, ethics, and compliance programmes or measures for preventing and detecting the bribery of foreign public officials in their international business transactions** (hereinafter "foreign bribery"), and to business organisations and professional associations, which play an essential role in assisting companies in these efforts." (p. 13)
2010	16	Fighting Corruption in the Supply Chain: A Guide for Customers and Suppliers (first edition)	UN (Global Compact)	Companies	"**For all companies, fighting corruption in the supply chain must be part of a larger anti-corruption programme that addresses corruption risks throughout the firm**." (p. 10)
	17	World Bank Group Integrity Compliance Guidelines	World Bank Group	Companies	"Going forward the establishment (or improvement) and implementation of **an integrity compliance program satisfactory to the WBG will be a principal condition to ending a debarment (or conditional non-debarment); or in the case of some existing debarments, early termination of the debarment."** (p. 1) "11. collective action: Where appropriate –

					especially for SMEs and other entities without well-established Programs, and for those larger corporate entities with established Programs, trade associations and similar organizations acting on a voluntary basis – endeavor to engage with business organizations, industry groups, professional associations and civil society organizations to encourage and assist other entities to develop programs aimed at preventing Misconduct." (p. 4)
2011	18	Business Against Corruption: A Framework for Action	UN (Global Compact), TI, International Business Leaders Forum	Companies	**Similar 2015 version**, adding "Companies may learn not only from their own actions, but also from the actions of others. Thus, companies can move together as a group, adopting successes and avoiding programmes that did not work for others." (p. 13)
	19	OECD Guidelines for Multinational Enterprises	OECD	Companies	"VII. Combating Bribery, Bribe Solicitation and Extortion […] In particular, **enterprises should**: […] 2. **Develop and adopt adequate internal controls, ethics and compliance programmes or measures for preventing and detecting bribery**, developed on the basis of a risk assessment addressing the individual circumstances of an enterprise, in particular the bribery risks facing the enterprise (such as its geographical and industrial sector of operation)." (p. 45)
	20	ICC Rules on Combating Corruption (fifth edition of ICC Rules)	ICC	Companies	"Article 10 Elements of a Corporate Compliance Programme **Each Enterprise should implement an efficient Corporate Compliance Programme** (i) reflecting these Rules, (ii) based on the results of a periodically conducted assessment of the risks faced in the Enterprise's business environment, (iii) adapted to the Enterprise's particular circumstances and (iv) with the aim of preventing and detecting Corruption and of promoting a culture of integrity in the Enterprise." (p. 11)
	21	World Bank Group Sanctioning Guidelines	World Bank Group	Companies	"Mitigating Factors […] Effective compliance program: Establishment or improvement, and implementation of a corporate compliance program. The timing, scope and quality of the action may indicate the degree to which it reflects genuine remorse and intention to reform, or a calculated step to reduce the severity of the sentence." (p. 4-5)

2012	22	Assurance Framework for Corporate Anti-bribery Programs	TI	Companies	"**Good practice now demands that enterprises develop comprehensive programmes to counter bribery in their business dealings which are monitored and improved on a continual basis**. This voluntary Assurance Framework aims to provide a standardized process that will help enterprises to design robust anti-bribery programmes." (p. 2)
	23	Global Compact for the 10th Principle Reports: Corporate Sustainability with Integrity - Organizational Change to Collective Action	UN (Global Compact)	Collective action	"To combat this problem, the Global Compact promotes the implementation of rigorous anti-corruption measures though organization change at the company level and collective action at the country level. First, **companies are asked to integrate anti-corruption and compliance measures into their business strategies and operations.** Companies develop their own code of conduct, including the implementation of a zero tolerance policy and a range of rules and regulations concerning gifts, political contributions, charities and travel. To apply these policies, companies implement a range of actions, including the establishment of anonymous hotlines, employee training, supply chain management, risk assessment and disciplinary measures. Second, companies are asked to take part in collective action, multi-stakeholder dialogue, and integrity or compliance pacts with industry peers." (p. 4) "**Certification of the Integrity Programme** […] Third, it **helps the company to continuously improve its compliance programme** based upon the recommendations given by external experts." (p. 6)
	24	Corruption Prevention to Foster Small and Medium-Sized Enterprise Development (second version)	UN (UNIDO and UNODC)	Collective action	"Even though internal measures are usually implemented more easily and quickly in SMEs than in large companies, **internal codes of conduct and compliance programmes alone are in many cases not helpful for SMEs**. SMEs are usually not motivated to adhere to ethical business practices (as they face limited pressure from stakeholders) or they lack the financial resources or the market power to enforce their zero-tolerance policies." (p. vii)
2013	25	An Anti-Corruption Ethics and Compliance Programme for Business: A Practical	UN (ONODC)	Companies	"It is now generally accepted that businesses have a responsibility to act as good corporate citizens. This tenet is increasingly complemented with evidence and **understanding among companies that fighting corruption makes good business sense and that a well-executed anti-**

		Guide			**corruption ethics and compliance programme yields greater value over time**." (p. 1)
	26	Anti-Corruption Ethics and Compliance Handbook for Business	OECD, UN (UNODC), World Bank	Companies	"The idea for this handbook began with G20 governments looking for ways to practically implement the 2010 G20 Anti-Corruption Action Plan. This Plan recognises the integral role the private sector plays in the fight against corruption and calls for greater public-private partnership in this effort. […] **The OECD, UNODC, and World Bank hope this handbook will be a useful resource not only for companies headquartered in G20 countries, but for all companies that recognise the need for developing and implementing robust anti-corruption ethics and compliance programmes**." (p. 3)
	27	A Resource Guide on State Measures for Strengthening Corporate Integrity	UN (UNODC)	Governments	"**The implementation of a meaningful and effective anti-corruption programme for business is primarily a private sector function and responsibility**. Anti-corruption measures are an investment, and like other business investments, they must compete with other demands for scarce resources based on perceived risks and benefits. **States can help to shape these corporate investment decisions through a combination of enforcement sanctions and good practice incentives**." (p. 2)
	28	Business Principles for Countering Bribery: A Multi-stakeholder Initiative led by Transparency International	TI	Companies	"The Business Principles The enterprise shall prohibit bribery in any form whether direct or indirect. **The enterprise shall commit to implementing a Programme to counter bribery. The Programme shall represent the enterprise's anti-bribery efforts including values, code of conduct, detailed policies and procedures, risk management, internal and external communication, training and guidance, internal controls, oversight, monitoring and assurance**." (p. 5)
	29	Corporate Integrity: Incentives for Corporate Integrity in Accordance with the United Nations Convention	UN (UNODC)	Governments	"The most frequently used sanction is a fine, which is sometimes characterized as criminal, sometimes as non-criminal and sometimes as a hybrid. Other sanctions include exclusion from contracting with the government (for example public procurement, aid procurement and export credit financing), forfeiture, confiscation, restitution, debarment or closing down of legal entities**. In addition, states may wish to consider non-monetary sanctions**

		Against Corruption – A Report			**available in some jurisdictions, such as** withdrawal of certain advantages, suspension of certain rights, prohibition of certain activities, publication of the judgement, the appointment of a trustee, **the requirement to establish an effective internal compliance programme** and the direct regulation of corporate structures." (p. 30)
2014	30	G20 ACWG Action Plan (2015-2016)	G20 (ACWG)	Governments	"G20 countries recognise that governments cannot fight corruption alone, and the private sector is an essential partner in helping us to achieve our anti-corruption goals. **G20 countries commit to continuing to work with the private sector and civil society to combat corruption**, including by developing anti-corruption education and training for business, with a particular focus on SMEs, and by examining best practices for **encouraging businesses to implement robust compliance programs and self-report breaches of corruption laws**." (p. 2)
	31	PACI Global Principles for Countering Corruption: Application & General Terms of Partnership *Vide supra note 242*	PACI	Companies	"Annex I: Guidance on Compliance While a vast majority of firms have sophisticated compliance programmes in place, the following section outlines the minimum requirements that businesses should meet when designing and implementing an effective anti-corruption programme. An effective programme comprises the entirety of an enterprise's anti-corruption efforts, specifically including its code of ethics, policies and procedures, risk assessment, internal and external communication, training, guidance, internal controls, monitoring and oversight. **The programme should cover corruption in all its forms**." (p. 8)
2015	32	Business Principles for Countering Bribery: Commentary	TI	Companies	"Strictly speaking, there should be only one Principle, prohibition of bribery, as the second Principle is a means to achieve the no-bribery Principle. **However, committing to implement an anti-bribery Programme is of such importance that it has been made a Principle**. The intent of this Principle is that the Board and senior management recognise that countering bribery requires a strategic approach to ensuring a culture of integrity in the enterprise and, based on continuing risk assessment, the design, implementation and maintenance of a range of policies and procedures to prevent and counter bribery." (p. 5)
	33	G20 High-Level Principles on	G20	Companies	"The private sector is an essential partner of governments in the fight against corruption, and its commitment to transparency and

		Private Sector Transparency and Integrity			integrity plays an integral role in achieving anticorruption goals. […] **The measures listed in this document are suggested general elements for developing or enhancing effective internal controls and ethics and compliance programs**. There is no 'one size fits all' approach. Emphasis on specific elements will vary from one business to another depending on, among other factors, the particular risks engendered by the business. **A business may wish to consider seeking advice from compliance or other professionals to learn more about what kind of internal controls and ethics and compliance program is most appropriate for its business and the jurisdictions where it operates."** (p. 1)
	34	G20 Principles for Promoting Integrity in Public Procurement	G20	Governments	"**8. G20 countries should foster a culture of integrity in public procurement among suppliers by:** **8.1 Encouraging supplier efforts to develop internal corporate controls, and compliance measures, including competition and anti-corruption programs and looking at ways in which due recognition could be given to suppliers that have effective controls, measures and programs in place**." (p. 3)
	35	G20/OECD Principles of Corporate Governance (third edition)	G20 and OECD	Companies	"**Compliance programmes should also extend to subsidiaries and where possible to third parties, such as agents and other intermediaries, consultants, representatives, distributors, contractors and suppliers, consortia, and joint venture partners**." (p. 26)
	36	Recommendation of the Council on Public Procurement (second edition)	OECD	Governments	"III. RECOMMENDS that Adherents preserve the integrity of the public procurement system through general standards and procurement-specific safeguards. To this end, **Adherents should: […] iv) Develop requirements for internal controls, compliance measures and anticorruption programmes for suppliers, including appropriate monitoring**." (p. 6-7)
2016	37	Ministerial Declaration – The Fight Against Foreign Bribery: Towards a New Era of	OECD	Companies and Governments	"**Invite the business community** to increase its **co-operation with governments in** the fight against foreign bribery and corruption and **encourage wide implementation of the OECD 2010 Good Practice Guidance on Internal Controls, Ethics and Compliance** developed by the Working Group." (p. 5) "**Encourage ongoing international efforts to identify and promote good practice in**

		Enforcement			**prevention of foreign bribery and corruption, which may include promoting the use of anti-corruption compliance measures, codes of conduct, and appropriate safeguards in public procurement processes such as those related to organising major international events**." (p. 6)
	38	Fighting Corruption in The Supply Chain: A Guide for Customers and Suppliers (second edition)	UN (Global Compact)	Companies	Concerning compliance programs, this version is the same as the 2010 one.
2017	39	G20 High-Level Principles on the Liability of Legal Persons for Corruption	G20	Governments	"Principle 14: Concrete incentives should be considered to foster effective compliance by businesses. While government enforcement of anti-corruption laws against legal persons is an essential component of an effective corporate liability regime, the private sector also has a key role in the development and implementation of effective compliance mechanisms within businesses. **Countries may therefore take into consideration, as appropriate, the existence of corporate anticorruption ethics and compliance programmes or measures in public procurement decisions or other processes to grant public benefits such as export credits. Moreover, efforts made by businesses to develop and implement effective anti-corruption internal controls, ethics and compliances programmes or measures, as well as voluntary self-reporting and cooperation by businesses with law enforcement may also, where appropriate and consistent with a country's legal system, be taken into consideration in legal proceedings, for example, as a potential mitigating factor or as a defence**." (p. 7)
2018	40	Stories of Change: Better Business by Preventing Corruption	TI	Companies	"Companies increasingly recognise that integrity is good for business. Yet bribery and corruption persist. Large-scale corporate scandals show that much remains to be done to tackle corruption in the business sector. **Based on four case studies, this paper shows how Transparency International is supporting companies worldwide to develop anti-corruption systems, which can help prevent corruption and boost**

					business, so that more companies reap the benefits of high integrity and transparency standards." (p. 2)
	41	The Future of Trust and Integrity	WEF (PACI)	Companies	"Integrating comprehensive compliance programmes into a business's operational processes can promote compliant and successful operations while mitigating corruption risks." (p. 16)
2019	42	Best Practices to Prevent and Combat Corruption	OAS (MESICIC)	Governments	Mexico: Register of Business Integrity "*En el marco de este Padrón de Integridad, se otorgará un Distintivo de Integridad Empresarial, el cual reconocerá a las empresas con buenas prácticas anticorrupción [...]Además, en colaboración con otras dependencias, organismos internacionales, organizaciones empresariales y la academia, se contará con mecanismos de evaluación empresarial para la supervisión, elaboración de formularios y herramientas, así como para el acompañamiento y asesoría para las empresas, especialmente para que las más pequeñas cuenten con protocolos de integridad, a fin de que puedan ser incorporadas al desarrollo de mejores prácticas y de una cultura de integridad*." (p. 1)
	43	G20 Compendium of Good Practices for Promoting Integrity and Transparency in Infrastructure Development	G20	Governments	"Assuring the integrity of bidding companies, for instance by: […] - **Establishing integrity programmes and guidelines for the private sector, where appropriate. […] Establishing close cooperation with business sector to implement the business integrity development programme**." (p. 9)
2020	44	Agenda for Business Integrity: Collective Action	WEF (Global Future Council on Transparency and Anti-corruption)	Collective action	"Capacity-Building Initiatives: **Companies jointly share their know-how, resources and tools from their compliance programmes**, and with the help of their compliance practitioners, to offer concrete capacity building and training opportunities for other companies that are part (or not) of their supply and value chains, in particular SMEs, as well as for public officials and organizations, and other practitioners from civil society organizations. The aim of these initiatives is to help create or enhance compliance systems and tools in smaller and/or less resourceful organizations." (p. 4)

	45	Connecting the Business and Human Rights and the Anti-Corruption Agendas: Report of the Working Group on the Issue of Human Rights and Transnational Corporations and other Business Enterprises	UN	Companies and Governments	"**Companies should include human rights due diligence and implementation of the Guiding Principles as part of a larger programme of compliance, sustainability and responsible business conduct**. This may involve integrating anti-corruption with human rights due diligence processes; at a minimum, it should involve alignment and recognition that both are key to responsible and sustainable business conduct. While there is no one size-fits-all solution, the responsibility to respect human rights is the baseline requirement." (p. 22) "**States should: […] (c) Introduce regulations that require human rights due diligence by business enterprises in line with the Guiding Principles, and provide guidance clarifying the connection between corruption and human rights risks and impacts**;" (p. 22)
	46	Ethics and Integrity Beyond Compliance: Compliance Agenda for Business Integrity	WEF (Global Future Council on Transparency and Anti-corruption)	Companies	"**Increasing business complexity and regulation has driven the evolution and strengthening of compliance programmes, together with the growth in influence and prominence of the function itself, and today there is a clear consensus about the need for such programmes and their key components**." (p. 4)

	47	Good Intentions, Bad Outcomes? How Organizations Can Make the Leap from Box-Ticking Compliance to Building a Culture of Integrity	WEF (Global Future Council on Transparency and Anti-corruption)	Companies	"Despite the widespread adoption of anti-corruption compliance programmes with all these features, however, corporate corruption and integrity scandals are still common. Therefore, **the dominant approach to anti-corruption compliance, whereby effort is focused on identifying and sanctioning individuals with unethical intent, is becoming less credible in the face of evidence that systemic corruption and fraud have taken root in a range of large multinational organizations that had established compliance systems**. […] **Compliance teams have gained visibility and resources. But, in many cases, the compliance team has come to be seen as an internalized law enforcement body that responds to external pressure from government regulators and the public.** Case in point, it tends to be staffed by lawyers – particularly, former prosecutors. This perception can have negative, unintended consequences and might even encourage employees to rationalize and justify unethical behaviour." (p. 3) "This paper shares key concepts that might help to advance beyond tick-box compliance programmes, towards true cultures of integrity in corporations." (p. 4)
2021	48	G20 ACWG Action Plan (2022-2024)	G20 (ACWG)	Governments	"**We will continue to encourage and support efforts by the private sector to strengthen effective internal controls and anti-corruption ethics and compliance programmes**, including for small and medium sized enterprises (SMEs) and the non-financial professional services sector." (p. 6) "In particular, the ACWG will: […] **Promote good practices in business integrity and anti-corruption ethics and compliance programmes**, covering issues such as maintenance of books and records, financial statement disclosures, accounting and auditing, and taking appropriate remedial steps to address wrongdoing." (p. 7)
	49	Recommendation of the Council for Further Combating Bribery of Foreign Public Officials in International Business	OECD	Companies and Governments	"D. Incentives for compliance Member countries should: i. **encourage their government agencies to consider, where international business transactions are concerned and as appropriate, internal controls, ethics and compliance programmes or measures for the purpose of preventing and detecting foreign bribery in their decisions to grant**

		Transactions (second edition)			**public advantages, including public subsidies, licences, public procurement contracts, contracts funded by official development assistance, and officially supported export credits;** ii. where member countries implement measures to incentivise enterprises to develop such compliance programmes or measures, provide training and guidance to their relevant government agencies, on how internal controls, ethics and compliance programmes or measures are taken into consideration in government agencies' decision-making processes, and ensure such guidance is publicised and easily accessible for companies; iii. **encourage law enforcement authorities, in the context of enforcement of the foreign bribery and related offences, to consider implementing measures to incentivise companies to develop effective internal controls, ethics, and compliance programmes or measures, including as a potential mitigating factor.** […] iv. where member countries implement measures to incentivise companies to develop such compliance programmes or measures, ensure that competent authorities consider providing training and guidance on assessing the adequacy and effectiveness of internal controls, ethics and compliance programmes or measures for the purpose of preventing and detecting foreign bribery, as well as on how such programmes or measures are taken into consideration in the context of foreign bribery enforcement, and ensure such information or guidance is publicised and easily accessible for companies, where appropriate." (p. 16) "Public Advantages, including Public Procurement XXIV. RECOMMENDS that: […] iii. **where appropriate and to the extent possible, in making such decisions on suspension and debarment, member countries take into account, as mitigating factors, remedial measures developed by companies to address specific foreign bribery risks, as well as any gaps in their existing internal controls, ethics, and compliance programmes or measures;**

					iv. member countries provide guidance and training to relevant government agencies on such suspension and debarment measures applicable to companies determined to have bribed foreign public officials and on remedial measures which may be adopted by companies, including internal controls, ethics and compliance programmes or measures, which may be taken into consideration;" (p. 16) "Annex II: **Good practice guidance on internal controls, ethics, and compliance**" (p. 20) – **Similar to the 2009 version**.
2023	50	Best Practices to Prevent and Combat Corruption	OAS (MESICIC)	Governments	*Paraguay: Sello Integridad* "*El Sello es un programa de incentivos a la integridad, el cual contribuye a fomentar programas de integridad en el sector empresarial paraguayo (los programas de integridad están compuestos de medidas y acciones para prevenir, detectar y remediar actos de corrupción y fraude, así como de acciones para promover una cultura organizacional de integridad), y concientizar a las empresas sobre su rol referente a la prevención de la corrupción y el impacto de ese tipo de hechos en la economía y el clima de negocios; buscándose, además, la difusión de buenas prácticas de integridade.*" (p. 1)
	51	OECD Guidelines for Multinational Enterprises on Responsible Business Conduct	OECD	Companies	"VII. Combating Bribery and Other Forms of Corruption […] In particular, **enterprises should**: […] 2. **Develop and adopt adequate internal controls, ethics and compliance programmes or measures for preventing, detecting, and addressing bribery and other forms of corruption**, developed on the basis of a risk-based assessment, taking into account the individual circumstances of an enterprise, in particular the risks factors related to bribery and other forms of corruption (including, *inter alia* its geographical and industrial sector of operation, other responsible business conduct issues, the regulatory environment, the type of business relationships, transactions with foreign governments, and use of third parties)." (p. 41)
	52	Recommendation of the Council on Principles of Corporate Governance (fourth	OECD	Companies	Concerning compliance programs, this version is the same as the 2015 one.

		edition)			

* Documents published in the same year are listed in alphabetical order.

NOTE

I would like to thank Professor Mariana Pargendler and Professor Marta Machado for their insightful comments. I also appreciate the support provided by the Fundação Getulio Vargas through the Mario Henrique Simonsen Scholarship for Teaching and Research; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; and São Paulo Research Foundation (FAPESP) - process 2021/13143-0.

PERFORMANCE PUBLIC POLICY: CHINA IN COVID-19

Xiang Gao*

Abstract: COVID-19 created a worldwide public policy vacuum due to the lack of scientific knowledge concerning effective disease control and vaccine. In this policy vacuum governments often sought to display themselves effective protectors of the public's health and safety despite a less than effective or complete policy performance. From this perspective is useful to compare actual policy outcomes as well as analyse the symbolic performance in public policy. This article presents an analytical framework of performance public policy with three elements, including constructing policy achievement, providing political goods, and establishing 'normative' commitment in both domestic politics and foreign policy. The research argues that Chinese government and CCP have been able to maintain a relatively high degree of social coherence and domestic support during the pandemic by employing performance public policy, a combination of materials strength, political propaganda, nationalist discourse, and assertive foreign policy.

Keywords: COVID-19; Performance Public Policy; Chinese Politics; Chinese Foreign Policy; Global Governance

* School of Humanities, Arts, and Social Sciences, University of New England, Australia.

Table of Contents

INTRODUCTION

The COVID-19 pandemic devastated global public health and deeply impacted the world economy. Unsurprisingly given the serious economic, social and political consequences of the pandemic, differing government responses have been highly politicised. Various quarantine and infection case tracking methods caused concern over the appropriateness of state power intruding into previously private spheres and on individual liberties. The usage of face masks, social distancing rules, and intra-state travel restrictions have aroused passionate debate over public health restrictions and human rights. In many states immigration policies, often stoked by racialist rhetoric, became more controversial and less humane.

Wuhan, the capital of Hubei Province, China experienced the first major outbreak of the novel coronavirus. China's official news reported that the earliest infections were identified on 8 December 2019. On 31 December the World Health Organization (WHO) was informed about the emergence of a 'pneumonia' of an unknown aetiology, which was later named COVID-19. The subsequent transmission of COVID-19 outside across the globe has resulted in over 700 million confirmed cases and approximately 7,000,000 deaths.[1] In addition to human suffering and economic misery, the outbreak has also triggered political tensions and the deterioration of bilateral relationships between China and many other countries, who criticized it for failing to live up its international responsibilities by failing to regulate activities that likely led to the initial infections, such as the wildlife trade, or its failure to limit its initial spread by mishandling health responses and inadequately informing international health authorities. Because of these shortcomings, American President Trump demanded compensation from China, a sentiment echoed by politicians and news outlets in Britain, France, Germany, and Australia. In particular, Chinese-Australian relations significantly deteriorated after Australia pushed for an international investigation into the COVID-19 outbreak. These international disputes have been accompanied by increasing racist or discriminatory animus directed towards people of Chinese or Asian descent in many states.

Nevertheless, in spite of the mismanagement and lack of transparency, the Chinese government's 'performance' was considered rather successful in the eyes for many domestic and international audiences from the end of January 2020 to late 2022, when the Government abruptly abandoned its "zero-covid" policy in the face of domestic fatigue and the increasing damage to the economy. After the initial missteps, Chinese authorities implemented strict measures to limit transmission and treat infected individuals. Medical personnel from across the country were sent to support Wuhan's hospitals. The government built two special COVID-19 hospitals in two weeks. Strict quarantine protocols were enforced: travel was restricted across regions, local residents could shop groceries only once a fortnight with a pass limited to one person for each household, and face masks were required at all times in public. By the middle of March 2020, new infections were near zero.

Prior to the wide availability of vaccines, this highly contagious disease presented difficult public policy challenges. The initial lack of an effective vaccine or treatment, high transmission rates and lack of knowledge concerning infection transmission, the difficulty of prevention, the structural stresses on health care systems,

[1] Worldometer, *Covid-19 Coronavirus Pandemic*, https://www.worldometers.info/coronavirus/.

and economic dislocation caused by the pandemic initially created a public policy vacuum across the world. This policy vacuum, with attendant issues relating to governmental performance, competence and legitimacy, is especially problematic for the Chinese Government. As the national economic growth lessens and the communist ideology becomes less relevant among the younger population, the effective management of issues of great public concern, such as COVID-19 by the Chinese Communist Party (CCP) can have important implications for regime legitimacy and security. As President Xi Jinping put it, the coronavirus is a 'crisis' and a 'major exam' for the Party leadership.

This article examines Chinese government's responses to COVID-19. It argues that by adopting a stylized 'performantive public policy', the Chinese government and CCP enhanced its domestic legitimacy through its manipulation of symbols and rhetoric coupled with the use of political "performance" or show in addressing the pandemic. First, the Government has generated a convincing policy performance for the Chinese domestic public by using a "result-oriented" pandemic policy. This featured strict quarantines, effective case tracking and aggressive research into treatment and vaccines joined with the wide dissemination of populist imagery featuring top political leadership actively engaged in combating the disease. Second, the Government fostered and re-emphasised nationalist and anti-West political discourses after the disease outbreak. This has the effect of diverting domestic attention by reemphasising the differentiation between 'us' and the 'others' among the Chinese public. Third, the Government pursued a more assertive foreign policy. This policy has been framed as a necessary response to the 'enemy outside' and is reinforced the assumed 'patriotic duty' of all Chinese people both at home and abroad to support the state.

I. PERFORMANCE, LEGITIMACY, AND PUBLIC POLICY

Weber argues that governments tend to legitimatize their rule on three main grounds: charismatic leadership, traditional leadership where rule is accepted because of religion, culture, and lineage; and rational-legal authority which is built upon a set of political institutions and bureaucratic procedures. Weber does not include regime performance as a source of legitimacy. Nevertheless, good socioeconomic performance generally enhances regime legitimacy. Political support is often associated with 'output effectiveness', including economic success and a high level of domestic satisfaction that people's needs are met on a day-to-day basis.[2] Examining the communist regimes in the 1980s, White points out that successful socioeconomic performance was essential for the legitimacy of the Soviet Union and other Eastern European communist states which otherwise lacked institutional and procedural based legitimacy.[3] The implication of this is notion is that when continuous economic growth cannot be achieved, the political management of the economic slowdown becomes crucial for stability and legitimacy.[4] Indeed, the rise of the Asian 'tigers' not only confirms the relationship between socioeconomic performance and regime durability, but also draws

[2] Ronald Rogowski, *Rational Legitimacy: A Theory of Political Support* (New Jersey: Princeton University Press, 1974), pp. 7-19.
[3] Stephen White, 'Economic performance and communist legitimacy', *World Politics* 38(3), (1986), pp. 462-482.
[4] Ibid.

attention to state capacity and autonomy, which can be used to effectively pursue developmental goals.[5]

Research on this broader notion of performance-based policymaking (and the impacts the success or non-success of particular policy programmes can have on regime legitimacy) has largely focused on the policy effectiveness, regime legitimacy, capacity building and state transformation. Authoritarian regimes and the post-conflict states in post-Cold War era have provided much empirical evidence for this scholarship. For example, Vietnam adopted *Doi Moi* policy in 1986, an economic reform aiming to establish a market-oriented economy. The consequent higher socioeconomic performance reinvigorated the legitimacy for the Communist Party of Vietnam (CPV), replacing traditional sources of legitimacy, such as socialist ideals and Ho Chi Minh's charismatic authority.[6] When positive socioeconomic performance could not be sustained, the CPV resorted to nationalism, evident in Vietnamese disputes with China in the South China Sea, to supplement its performance-based legitimacy.[7] Suharto's Indonesia is another example of performance-based legitimacy in an authoritarian state. The Suharto regime sought to build strong state capacity featuring improved civil service in the context of stable economic growth. This approach, coupled with a reliance on patronage generated significant domestic support until Asian Financial Crisis interrupted economic growth leaving the regime vulnerable.[8] Soest and Grauvogel have argued that performance legitimacy can be derived from the delivery of public goods such as security, education or health care.[9] Examining the post-conflict states, such as Afghanistan and South Sudan, Dagher argues that performance legitimacy is earned by state and non-state actors when they deliver public goods, services and welfare that are urgently associated with the daily lives of citizens.[10] This output based legitimacy, bolstered by institutional capacity building, is particularly important where the public has limited experience with liberal democratic culture.[11]

While state performance has been mostly defined in material terms, recent scholarship looks beyond the socioeconomic outputs and has increasingly focused on the ideational and normative criteria against which citizens evaluate state performance. From this perspective, a state's ability to deliver services and economic benefits does *not* necessarily lead to regime legitimacy and stability. First, many intervening variables interrupt this seemingly straightforward causal relationship between performance outputs and legitimacy, such as citizens' changing expectations towards government, the equality of public goods distribution, management of service delivery,

---

[5] Adam Przeworski and Fernando Limongi, 'Political regimes and economic growth', *Journal of Economic Perspectives* 7(23), (1993), pp. 51-69.
[6] Hong Hiep Le, 'Performance-based legitimacy: The case of the Communist Party of Vietnam and "Doi Moi"', *Contemporary Southeast Asia* 34(2), (2012), pp. 145-172.
[7] Ibid.
[8] Marcus Mietzner, 'Authoritarian elections, state capacity, and performance legitimacy: Phases of regime consolidation and decline in Suharto's Indonesia', *International Political Science Review* 39(1), (2018), pp. 83-96.
[9] Christian von Soest and Julia Grauvogel, 'Identity, procedures and performance: how authoritarian regimes legitimize their rule', *Contemporary Politics* 23(3), (2017), pp. 287-305.
[10] Ruby Dagher, 'Legitimacy and post-conflict state-building: The undervalued role of performance legitimacy', *Conflict, Security & Development* 18(2), (2018), pp. 85-111.
[11] Ibid.

and the attribution process.[12] Second, other than the widely accepted socioeconomic indicators, state performance can also be measured in non-material forms. More specifically, the ability to provide common political goods, such as civil and political rights, law and order, the absence of corruption, government and political leaders' responsiveness, national identity, and shared values, is another set of criteria to gauge governmental effectiveness. Providing 'order, protection, safety, trust, and the conditions of cooperation' is often times sufficient to secure legitimacy.[13] State performance is also assessed on the basis of moral principles and normative commitments, creating demanding requirements for state performance and legitimacy. This 'moral performance' while arising from internal domestic morality/ethics also encompasses international normative obligations and rhetoric. The internal ethics generates political support from citizens over whom state power is exercised;[14] while a commitment to international norms secures legitimisation from the global normative community.

II. PERFORMANCE AND CHINESE PUBLIC POLICY

Scholarship on Chinese politics has tended to attribute regime stability and legitimacy to the socioeconomic performance that the Chinese government and CCP have been able to generate since 1978. Indeed, many observers have argued that the discussion of legitimacy can be simplified to an assessment on 'governance'.[15] More specifically, this performance-based legitimacy is measured by Chinese government's ability to promote and sustain economic growth and social stability through solid governance policies and political institutions.[16] The reform and ownership diversification of state-owned enterprises and other smaller private enterprises, the transition from centrally-planned to a more market-oriented economy, and the development of trade and foreign investment, have created a dynamic Chinese economy over the past 40 years. According to World Bank, China's annual GDP growth has averaged close to 10 percent, and over 850 million people have been raised from poverty since 1978.[17] To secure social stability in the face on this economic and social transformation, the Chinese government has adopted two types of policies. On the one hand, it implemented programmes to assist vulnerable groups, such as the workers who lost their employment due to state-owned enterprise reform and rural students who cannot afford education. On the other hand, it has also adopted repressive policies towards political dissidents, democracy advocates and human rights activists while actively censuring or disrupting potential sources of dissenting public speech or action.[18] The resulting sustained economic achievement and social stability have reaped the Party political goodwill and capital, while providing support for CCP's

---

[12] Clair Mcloughlin, 'When does service delivery improve the legitimacy of a fragile or conflict-affected state?', *Governance: An International Journal of Policy, Administration, and Institutions* 28(3), (2015), pp. 341–356.
[13] Bernard Williams, *Realism and Moralism in Political Theory* (New Jersey: Princeton University Press, 2005), p. 3.
[14] Edward Hall, 'Bernard Williams and the basic legitimation demand: A defence', *Political Studies* 63, (2015), pp. 466-480.
[15] Yucaho Zhu. '"Performance legitimacy" and China's political adaptation strategy', *Journal of Chinese Political Science* 16, (2011), pp. 123-140.
[16] Hongxing Yang and Dingxin Zhao, 'Performance legitimacy, state autonomy and China's economic miracle', *Journal of Contemporary China* 24(91), (2015), pp. 64-82.
[17] World Bank, 'The World Bank in China: Overview', https://www.worldbank.org/en/country/china/overview#1 (accessed 26 August 2020).
[18] Zhu, '"Performance legitimacy" and China's political adaptation strategy'.

leadership and authority.[19] Additionally, the government has skilfully rallied nationalist and patriotic sentiment to supplement its performance-based legitimacy.[20] Patriotic education campaigns have been launched to enhance national unity. These campaigns have emphasised China's historic victimhood, the West's 'ill intention' and containment policy towards China, and the 'patriotic duty' of all ethnic Chinese to support the PRC despite their citizenship.

Having recognised the salience of socioeconomic performance and nationalism in Chinese politics, recent research has broadened the notion of policy performance, while paying more attention to the symbolic meaning and normative interpretations of those factors which comprise policy performance in the eyes of policymakers and the public. Scholarship on the Communist Party rule in China has led to even more expansive notion of governmental "performance" to address the unique aspects of Chinese state, economy and society that have surfaced over the past two decades. For instance, Dickson writes: '…to the extent that the Chinese public regards the current regime as legitimate, it is primarily on the basis of performance legitimacy — specifically with regard to modernization, nationalism, and political stability.'[21] Zeng has expanded the definition to include the performance as an amalgam of "all government function" interwoven with ideology and nationalism.[22] This ideological-institutional approach suggests a close examination to CCP's ability to construct, shape, and institutionalise certain "pro-government" or "pro-Chinese 'subjective values and meanings' which are applied to evaluate China's policy performance.[23] From this perspective, a non-material, symbolic, or yet non-existing 'accomplishment' is as equally important as the concrete performance and material outcomes.[24] For example, the current Chinese leadership under Xi Jinping interprets and frames Chinese economic and political achievement since 1978 into a syllogism of national pride and collective satisfaction: the 'China Dream'- the rejuvenation of nation.[25] The Chinese government regularly showcases its policy performance (and legitimacy) through various public and cultural events, such as National Day parades, movies and songs featuring patriotic themes, and 'red' tourism.[26] In times of crisis, this symbolic policy performance, as described and explained through state-controlled media can shape citizens' perception of the crisis generating public support.[27] This performance public

---

[19] André Laliberté and Marc Lanteigne, eds. *The Chinese Party-State in the 21st Century: Adaptation and the Reinvention of Legitimacy* (London and New York: Routledge, 2008), p.15; Kerry Brown, *Contemporary China* (London: Red Global Press, 2019), p. 228.

[20] Michael Roskin, *Countries and Concepts: Politics, Geography, Culture* (New York: Pearson/Longman, 2009) p. 426; Philip P Pan, *Out of Mao's Shadow: The Struggle for the Soul of a New China* (New York: Simon & Schuster, 2008), p. 323.

[21] Bruce J Dickson, 'No "Jasmine" for China', *Current History* 110(737), (2011), pp. 211-216.

[22] Jinhan Zeng, The Chinese Communist Party's Capacity to Rule: Ideology, Legitimacy, and Party Cohesion (New York: Palgrave Macmillan, 2016), p. 68.

[23] Heike Holbig and Bruce Gilley, 'Reclaiming legitimacy in China', *Politics & Policy* 38(3), (2010), pp. 395-422.

[24] Seraphone F Maerz, 'The many faces of authoritarian persistence: A set-theory perspective on the survival strategies of authoritarian regimes', *Government and Opposition: An International Journal of Comparative Politics* 55, (2020), pp. 64-87.

[25] Tony Saich, *Governance and Politics of China* (London: Red Global Press, 2015), p. 76.

[26] Yih-Jye Hwang and Florian Schneider, 'Performance, Meaning, and Ideology in the Making of Legitimacy: The celebration of the People's Republic of China's sixty-year anniversary', *The Chian Review* 11(1), (2011), pp. 27-56.

[27] Jessica C Weiss and Allan Dafoe, 'Authoritarian audiences, rhetoric, and propaganda in international crises: Evidence from China', *International Studies Quarterly* 63(4), (2019), pp. 963-973.

policy can be effective and lower cost when compared to substantive policy responses, while still meeting public expectations and ensuring that state preferences remain essentially unchallenged.

III. PERFORMANCE PUBLIC POLICY

It is evident from the discussion above that the Chinese government constructs and epitomises its policy performance to shape public perception and enhance CCP's authority and legitimacy. In the face of Covid-19, ideational factors in Chinese public policy, more specifically the interpretation and symbolic meaning of policy performance that Chinese government and CCP used in order to promote a 'satisfactory' or paranematic policy outcome to the domestic public was particularly significant. An investigation of this phenomenon includes three overlapping and mutually reinforcing elements. First, there is a "Spin" (controlling or influencing communication in order to deliver a preferred message) element associated with the particular policy. Using "Spin" the Chinese government epitomizes policy outputs by presenting or constructing a material achievement (or non-achievement) in a favourable light. This is often accompanied by linking the output (or non-output) to be a direct result of competent populist or technocratic leadership. Second, there is the rhetorical and material provision of political and public goods. Along with tangible public goods which may be directed at a portion of the population, political goods which as not materially divisible, are also provided to optimize a positive public perception of government performance. These political goods, such things as national unity and pride, shared values, and strong leadership capacity are often boosted by anti-Western and nationalistic political discourse. Third, the Chinese government uses moralistic/ethical foreign policy tropes to demonstrate a 'moral commitment[s]' in its foreign policy to satisfy the domestic audience. Adopting an assertive foreign policy [both rhetorically and on the ground] and emphasising Western countries' 'wrongful conduct' against China, the Chinese government has fostered a domestically appealing moral 'high ground', that includes defending Chinese sovereignty and national interests, which in turn justifies and legitimates its foreign policy.

A. Constructing Policy Outcomes: 'For A Full Victory Against the Pandemic'[28]

The Covid-19 pandemic first broke out in Wuhan in December 2019. Initially ill-prepared, the Chinese government regrouped from early mistakes and essentially controlled transmission in about three months. Along with this substantial public health achievement, the government has also skilfully constructed the public policy outcomes during the pandemic; and presented them in a convincing manner to the domestic public which both lessened the real and perceived danger of the disease while enhancing its popularity and the positive perception of senior leadership. A positive, even heroic performance, coupled with cultivation of top political leaders' populist images as portrayed in the media enhanced national pride and secured additional domestic support, and by implication enhanced the legitimacy of the Chinese Government.

1. Spinning the Narrative of Achievement

[28] People's Daily, 'Jianjue daying hubei baoweizhan, wuhan baoweizhan' ['Determined to gain the victory of defending Hubei and Wuhan'] 15 March 2020, http://tyzx.people.cn/n1/2020/0316/c385048-31633362.html (accessed 28 August 2020).

a. Creating a Narrative

Chinese state media and propaganda apparatus attributed China's 'good performance' against the pandemic to the 'advantages of China's political system', CCP's leadership, Party member's dedication, and the sacrifice and efforts of all Chinese people. In the narrative, the public health policies deployed to battle the virus, highlighted by China's 'speed, scale, and efficiency', were lauded for their 'exemplary performance' by Chinese media.[29] This policy performance tended to be quantified and presented in a manner to exhibit superior performance. The declining number of new cases, increased hospital capacity, growing numbers of medical personnel and equipment, as well as increasing community compliance, were used to showcase the achievement, although further analysis suggests a more equivocal and nuanced evaluation. For example, from late April 2020 many major Chinese new outlets, such as People, Xinhua, and China National Radio, effusively celebrated the 'high recovery rate' (94.3%) and 'low fatality rate' (5.6%) of the Chinese COVID-19 patients.[30] Yet when compared to the global statistics, the results barely met the average global recovery rate of 95%, and remained below some other countries such as Australia and Germany.[31] Nevertheless, the media applauded the Party's leadership, (particularly the Party's central leadership), effective mass mobilisation, advanced scientific methods, and national unity for achieving this 'outstanding performance'.[32] The 2020 White Paper entitled '*Fighting COVID-19: China in action*' summarises China's 'strategic achievement' in the simple language of numbers: in a month, the rate of infection was contained; in two months, the daily reported cases, which had increased at the onset of the pandemic, fell to single digits; and in three months, a 'decisive victory' was secured in Wuhan City and Hubei Province.[33] This clear articulation of the positive government performance rallied political support. Indeed, a 2020 survey showed that 89 percent of citizens are satisfied with the government's information dissemination during the pandemic.[34]

In addition to domestic disease control, the Chinese government has also demonstrated and spun its superior policy performance by measures it took to protect Chinese citizens' health overseas. By late March 2020, the spread of virus within China was effectively under control; while new cases outside China had increased. The State Council Information Office in April 2020 revealed that President Xi Jinping had telephone communication with top political leaders of other countries, such as Britain,

---

[29] Minister of Foreign Affairs of PRC, 'Xi Jinping meets with visiting World Health Organization (WHO) Director-General Tedros Adhanom Ghebreyesus', 29 January 2020, https://www.fmprc.gov.cn/mfa_eng/zxxx_662805/t1737014.shtml (accessed 31 August 2020).

[30] For example, see Dong Changxi, 'Zhongguo xinguan feiyan zhiyulv weishenme zheme gao' ['Why is the recovery rate of coronavirus patients so high in China'], People.cn, 30 April 2020, http://health.people.com.cn/n1/2020/0430/c14739-31694518.html (accessed 2 September 2020).

[31] Worldometer, 'COVID-19 coronavirus pandemic', https://www.worldometers.info/coronavirus/#countries (accessed 1 September 2020).

[32] Pei Guangjiang, Huan Xiang, Xie Jianing and Rong Yi, ' "Zhongguo dajuan" jingdeqi lishi jianyan' ['Chinese response' can pass the test by history'], *People's Daily Online*, 8 June 2020, http://politics.people.com.cn/n1/2020/0608/c1001-31738044.html (accessed on 1 September 2020).

[33] China's State Council Information Office, 'Fighting COVID-19: China in action', *Xinhuanet*, 7 June 2020, http://www.xinhuanet.com/english/2020-06/07/c_139120424.htm?bsh_bid=5517099546 (accessed 2 September 2020).

[34] Cary Wu, 'How Chinese citizens view their government's coronavirus response', *The Conversation*, 5 June 2020, https://theconversation.com/how-chinese-citizens-view-their-governments-coronavirus-response-139176 (accessed 2 September 2020).

the United States, France, and Germany. Xi requested his counterparts to 'protect the health, safety and lawful rights' of Chinese citizens. It was reported that he had 'received positive responses.'[35] After many governments evacuated their citizens from China at the early stage of the pandemic, (an international embarrassment as it exhibited a lack of confidence in Chinese Covid-19 prophylactic and treatment measures) the Chinese government sought to change this narrative and 'image loss' by chartering flights to bring underaged Chinese overseas students back home. For example, on 2 April, the first chartered flight organised by Chinese embassy in UK took approximately 188 under-18 Chinese students from London to Jinan, Shandong Province.[36] Later, more chartered flights were arranged. These student repatriations were domestically acclaimed as evidence the government's care and compassion for vulnerable overseas students. Additionally, Chinese embassies provided more than 1 million 'health kits', containing face masks, anti-bacterial wipes, capsuled Chinese herbal medicine, and a COVID-19 educational pamphlet to those overseas Chinese students who could not return.[37] The gratitude of the student beneficiaries was widely publicised through various social media platforms such as WeChat. And major Chinese news outlets profusely praised the 'unity and deep love to the motherland' of younger generation, while noting that 'the motherland always supports her citizens overseas, and serving the people is the ultimate goal of Chinese government'.[38] This underpinned other discussions that accompanied reports on overseas students which questioned their patriotism and Chinese identity by suggesting that 'western' values and foreign influenced sensibilities had no room in the Chinese polity. Many commentators opined that the repatriated students could not 'positively contribute to the motherland's development' in the future because they had received a 'western education' from a young age.[39] This us/them positioning of China v. "the west" further underscored the significance and superiority of Chinese policy and its care for its citizens in China.

The government received criticism from the domestic public and international community for its lack of transparency and mishandling of cases, especially at the early stage of the pandemic. Facing the increasing number of fatal cases in Hubei Province in January 2020, the government, even while pursuing significant material health initiatives, nevertheless sought to present or reconstruct its past poor policy performance in more a favourable light. First, the central government distanced itself from any 'wrongful conduct' by assigning blame to the local governments for the mismanagement of quarantine and disease control measures. Numerous local government officials in multiple provinces, (e.g. Hubei, Hunan, Sichuan, Henan, Gansu,

---

[35] Gov.cn, 'Guowuyuan xinwuban jiu yiqing qijian zhongguo haiwai liuxue renyuan anquan wenti juxing fabuhui' ['State Council Information Office held news release regarding the safety of Chinese students overseas during the pandemic'], 2 April 2020, http://www.gov.cn/xinwen/2020-04/02/content_5498179.htm (accessed 4 September 2020).

[36] Sina Finance, 'Baoji jie xiao liuxuesheng huiguo!' ['Chartered flight taking young Chinese students overseas back home'], 1 April 2020, http://finance.sina.com.cn/wm/2020-04-01/doc-iimxxsth3034683.shtml (accessed 4 September 2020).

[37] Gov.cn, 'Guowuyuan xinwenban jiu zhongguo guanyu kangji yiqing de guoji hezuo qingkuang juxing fabuhui' ['State Council Information Office held news release regarding international cooperation to combat the pandemic'], 26 March 2020, http://www.gov.cn/xinwen/2020-03/26/content_5495712.htm#1 (accessed 4 September 2020).

[38] Gov.cn. 'Guowuyuan xinwuban jiu yiqing qijian zhongguo haiwai liuxue renyuan anquan wenti juxing fabuhui'.

[39] For example, see Sohu, 'Gaibugai jie waiguo de xiao liuxuesheng huiguo?' ['Should young overseas students be brought back home'], 26 March 2020, https://www.sohu.com/a/383272310_100214804 (accessed 4 September 2020).

Tianjin and Zhejiang) were disciplined for not strictly implementing quarantine rules.[40] The central government additionally encouraged the public to monitor local authorities' performance in disease control and report any misconduct through a State Council App launched in 2019.[41] Second, in response to the criticism related to the withholding of information on the disease, the Wuhan government corrected the COVID-19 case numbers and fatalities in April 2020. It explained that the 'oversight' of 1,290 undocumented deaths was largely due to the 'lack of hospital capacity', noting the correction was made to 'respect the history, the people, and those who lost their lives'.[42]

In addition to transparency issues, China's human rights violations during the pandemic also garnered international attention. Human rights advocacy groups were concerned with arbitrary detentions and restrictions on free speech, which have deepened with the COVID-19 lockdown. Domestic outrage also grew after news that Li Wenliang died from the virus in February, 2020. The Wuhan doctor who had voiced the public health concern over COVID-19 in November 2019 and subsequently received police reprimand and a formal written warning and censure for "publishing untrue statements about seven confirmed SARS cases at the Huanan Seafood Market."[43] His treatment with the authorities raised public and international concerns about the lack of free speech for those individuals concerned about the disease. In response to the opprobrium, the government proffered an alternative explanation which finessed the criticism and promoted its preferred interpretation of human rights. First, in the face of increasing netizens' praise for Li as a 'hero' and 'whistle-blower', the government redefined the nature of the matter. It stated that the police reprimand was a 'wrongful application of the rule of law' and should be revoked. Instead of a simple revocation and admonishment to the police, the government when further. It noted that Li was a 'true patriot' and titled him the honorific 'martyr' for his actions. Additionally, in a further effort to deflect criticism, the government emphasized that Li was an ophthalmologist and a CCP member. Because of his CCP membership any attempt to label him a 'whistle-blower', 'hero', and 'awakener' *against* the 'system' was an 'insult to Dr Li and his family' [i.e. he was in the system and so his protestations were simply a demonstration of the way the system is supposed to work].[44] Second, the government engaged in a broader discussion related to human rights, again promoting social-economic rights over civil-political rights. It argued that the right to life and health are the basic human rights, and as such should be prioritised in the global pandemic. More

[40] People.cn, 'Zhuyile! Zhexie ganbu yin yiqing fangkong buli deng bei yansu wenze' ['Attention! These cadres were held responsibility seriously for not effectively controlling the disease spread'], 29 January 2020, http://fanfu.people.com.cn/n1/2020/0129/c64371-31564153.html (accessed 7 September 2020).

[41] Gov.cn, 'Guowuyuan bangongting xiang quanshehui zhengji!' ['General Office of the State Council gathering information from the public'], 24 January 2020, http://www.gov.cn/hudong/2020-01/24/content_5472009.htm (accessed 7 September 2020).

[42] Xinhuanet, 'Wuhan dingzheng xinguan feiyan quezheng bingli he siwang shuju, ['Wuhan corrects the statistics of coronavirus cases and fatalities'], 18 April 2020, http://www.xinhuanet.com//mrdx/2020-04/18/c_138986696.htm (accessed 7 September 2020).

[43] Stephanie Hegarty, "The Chinese doctor who tried to warn others about coronavirus," 6 February 2020, BBC News, https://www.bbc.com/news/world-asia-china-51364382 (Accessed 12 March 2020).

[44] Ministry of Foreign Affairs of People's Republic of China, 'Guanyu shehua renquan wenti de gezhong miulun yi shishi zhenxiang' ['The fallacies and truth regarding human rights in China'], 2 July 2020, switzerlandemb.fmprc.gov.cn/web/zyxw/t1794112.shtml (accessed 4 September 2020.

specifically, 'equality among patients, protecting people's livelihood, open access to information, and rule of law' are the 'foundations' of human rights in China.[45]

b. Underpinning Populist Leadership

Populist politics is anti-establishment and anti-elitist. From this perspective, populists have a problematic relationship toward holding and maintaining power, as wielding power could make them a part of the 'corrupt elite', in opposition to 'the people'.[46] However, Chinese populism does not exhibit a genuinely anti-elitist nature, especially as it has been moderated by the government. Rather Chinese populism, consistent with other nationalist populism proffers an unmediated relationship between the 'paramount leader' and 'Chinese people' but does not exhibit a direct anti-elite animus. This is evident in the fact that positive policy outcomes are often delivered from the top political leadership directly to the public without intermediaries or institutional accoutrements. During the pandemic, President Xi Jinping's media appearances exhibited him as a caring, compassionate yet determined leader combating the pandemic together with the people. Xi's speech made in early February 2020 at the Beijing Disease Control Centre was widely publicised and repeatedly quoted, especially his statement that 'People's life, safety, and health are always the priority.' This image of a caring leader was similarly emphasised in news reports on his 10 March visit to Wuhan. According to reports, many officials and medical professionals were 'much encouraged' and 'moved by Xi's deep feeling for the people' both during and after the trip.[47]

Furthermore, all the success and 'policy achievement' against Covid-19 was generally attributed to Xi's personal dedication and leadership; while any undesirable performance outcome was 'due to the poor local responses'. Xi's personal association with the 'successful disease control' could be observed on all levels of political media. For example, in the White Paper, Xi's name and leadership are mentioned 49 times. The Paper emphasizes that CCP General Secretary Xi Jinping 'personally leads and coordinates the [anti-virus] action', which had given the people much 'confidence, strength, and guidance'.[48] In addition to disease control, Xi's personal commitment and interest in economic recovery, especially poverty alleviation, which was necessary in the face of the pandemic's economic dislocation, was also showcased. After March 2020 Xi carried out a series of highly publicized inspections of several economically less developed provinces, such as Shaanxi and Shanxi, and reiterated the importance of fostering local industries to benefit the public.[49]

---

[45] Zhang Yonghe, 'Zhongguo yiqing fangkong zhangxian renquan baozhang' ['Disease control in China shows the practice of protecting human rights'], *People.cn*, 20 March 2020, theory.people.com.cn/n1/2020/0320/c40531-31640521.html (accessed 4 September 2020).

[46] Cas Mudde and Cristóbal Rovira Kaltwasser, 'Populism' in *The Oxford Handbook of Political Ideologies* (Oxford: Oxford University Press, 2013), p. 503.

[47] People's Daily, 'Jianjue daying yiqing fangkong de renmin zhanzheng; ['Determined to win the People's War against the pandemic'], 11 February 2020, http://paper.people.com.cn/rmrb/html/2020-02/11/nw.D110000renmrb_20200211_1-01.htm (accessed 8 September 2020).

[48] China's State Council Information Office, 'Fighting COVID-19: China in action', *Chinese government White Paper*, Xinhuanet, 7 June 2020, http://www.xinhuanet.com/english/2020-06/07/c_139120424.htm?bsh_bid=5517099546 (accessed 2 September 2020).

[49] Cheng Yao, 'Liuci difang kaocha, Xi jinping guanzhu naxie zhongdian' ['What did Xi focus on in his six local inspection'], *Xihuanet*, 14 May 2020, www.xinhuanet.com/politics/xxjxs/2020-05/14/c_1125984071.htm (accessed 8 September 2020).

Of course, an unmediated relationship with the people can create problems if policy performance is perceived as less than adequate. Since assuming the 'paramount leadership,' due to his jettisoning of the collective leadership model used over the past 3 decades, Xi has had to shoulder more personal responsibly for policy outcomes. This can be both beneficial, underlining his personal concern and relationship with the population but it can also create problems. On one hand, Xi can enjoy the credit, esteem and popularity that comes with good policy performance; on the other hand, he can be held 'individually responsible' for any policy failure. For example, at the early period of virus spread (January 2020) many citizens were trying to buy face masks, but many online and retail outlets were sold out. Anxious netizens and local residents started to 'demand' President Xi, instead of local officials, to deliver them face masks.[50]

2. Providing Political Goods: 'The Great Spirit of China'[51]

Besides actually generating material achievement or creating a narrative that such achievement has occurred, an important aspect of state policy performance is the provision non-material benefits to the public. These non-material benefits (re)generate support for the policy choices undertaken and deepen the legitimacy of the government. This is particularly the case when there is high uncertainty and volatility such that it is difficult to assess any substantial government performance, a situation which occurred early in the COVID crisis. During the pandemic, the government exhibited this aspect of policy performance through the highlighting its provision of 'political goods', including law and order, national unity and pride, and shared values. For example, in February 2020, the Supreme Court, the Supreme Procuratorate, Ministry of Public Security, and Ministry of Justice jointly published an advisory opinion regarding legal punishment for offenses relating to the interference with disease control measures. Severe punishments were introduced for violence against medical professionals and police, producing and selling counterfeit medications, raising commodity prices, spreading 'rumours', and being uncooperative with quarantine measures. [52] Additionally, as an extension of Xi's highly popular anti-corruption campaign, many local political leaders, for example in Hubei, Guizhou, Guangxi, Jiangxi, Hunan, and Tianjin, were given Party discipline or criminal charges for inappropriate behaviour, abuse of power, and corruption during the pandemic.[53]

Second, government emphasised its effective treatment of COVID-19 patients, especially through the use of Chinese traditional medicine (TCM). Despite the lack of rigorous trial data on effectiveness of TCM, various TCM remedies such as herbal drink, were promoted and widely used as a treatment.[54] In an effort to publicize and share the 'Chinese experience' and 'Chinese solution', TCM remedies were also sent to other

---

[50] This is based on the author's observation in Jiangsu Province, China, in 23-30 January 2020.

[51] People.cn, 'Zai yiqing fangkong douzheng zhong zhangxian weida zhongguo jingshen' ['Demonstrating the great spirit of China in fighting the pandemic'], 7 April 2020, http://opinion.people.com.cn/n1/2020/0407/c1003-31663076.html (accessed 9 September 2020).

[52] Xu Juan, 'Wei yiqing fangkong zhulao fazhi diba' ['Build a strong legal 'dam' for disease control'], *People.cn,* 24 February 2020, opinion.people.com.cn/n1/2020/0224/c1003-31600409.html (accessed 10 September 2020).

[53] People.cn, 'Hubeisheng Huanggangshi chufen dangyuan ganbu 337 ren' ['337 Party cadres were discipline in Huanggang, Hubei Province'], 2 February 2020, fanfu.people.com.cn/n1/2020/0130/c64371-31565382.html (accessed 10 September 2020).

[54] David Cyranoski, 'China is promoting coronavirus treatments based on unproven traditional medicines', *Nature*, 6 May 2020, https://www.nature.com/articles/d41586-020-01284-x (accessed 2 September 2020).

countries, such as Thailand, Iran and Italy, as a form of international aid during the pandemic.[55] According to the 2020 White Paper, TCM was involved in treating 92% patients, and was proven useful in over 90% cases.[56] President Xi championed the use of TCM as 'a treasure of Chinese civilisation'.[57] The promotion of traditional medicine by the government clearly fostered a heightened level of national pride and confidence that in turn enhanced CCP's authority and legitimacy in uncertain times.

Third, Chinese state media described fighting COVID-19 a global-wide 'competition'. In this competition, a nation's 'material power' as well as 'mental strength'(which calls for the highest level of nation unity and patriotism), is put to test. The official governmental discourse in media and through governmental information releases placed this 'competition' in light of the national mythology as it related to the formation of Chinese nation, the historical 'hardship', and the 'heroic Chinese people' to rally political support.[58] Further as the early suppression policies proved more and more effective China 'won this competition' (for there are no infection within the country), the Chinese people have demonstrated the 'great spirit of China' to the world.

In addition to emphasizing Xi's 'paramount leadership' in this competitive fight against Covid-19, the political discourse during the pandemic skilfully superimposed and equalised the terms of 'Chinese people', 'Chinese nation', CCP, and People's Liberation Army. The policy achievement of overcoming the virus, is built upon the 'heroism of the whole Party, Army, and Chinese people from all ethnic groups,' and therefore the 'true patriotism' requires an individual to equally and unequivocally support to all these entities.

After transmission was brought under control in April 2020, there was an outpouring national pride and confidence. Chinese media has presented the government's response to COVID-19 as a living evidence of 'China dream': 'the sincere wish of a shared national destiny' and 'the great expectation of a strong and prosperous state' which tie 'all Chinese people' together.[59] Such nationalist pride and patriotism was exemplified by a six-hour documentary series 'Fighting Together' (tongxin zhanyi 同心战"疫"), documenting and celebrating this success was jointly made by the Publicity Department of CCP and China Central Television (CCTV). It aired in September 2020 to a mainly domestic audience. The documentary, grounded in historical fact but with considerable artistic license, pays tribute to Wuhan city where the pandemic first appeared, and praises the 'heroic deeds' of medical professionals, CCP party members, PLA, and ordinary Chinese people. President Xi Jinping's leadership, policy instructions and personal commitment echo through the documentary, even though Xi was criticised for his absence during the earlier days of the pandemic.[60]

---

55 Ibid.
56 China's State Council Information Office, 'Fighting COVID-19: China in action'.
57 Nectar Gan and Yong Xiong, 'Beijing is promoting traditional medicine as a "Chinese solution" to coronavirus. Not everyone is on board', *CNN*, 16 March 2020, https://edition.cnn.com/2020/03/14/asia/coronavirus-traditional-chinese-medicine-intl-hnk/index.html (accessed 2 September 2020).
58 People.cn. 'Zai yiqing fangkong douzheng zhong zhangxian weida zhongguo jingshen'.
59 People.cn, 'Zai yiqing fangkong douzheng zhong zhangxian weida zhongguo jingshen'.
60 Chris Buckley and Steven Lee Myers, 'Where's Xi? China's leader commands coronavirus fight from the safe heights', *New York Times*, 8 February 2020, https://www.nytimes.com/2020/02/08/world/asia/xi-coronavirus-china.html (accessed on 4 December 2020).

State media, Xinhua, described the documentary series as an important demonstration of President Xi's and CCP's leadership as well as the 'advantages of China's socialist political system'.[61] Around the same time, the Ministry of Culture and Tourism sponsored the opera entitled 'Angel's Diary', telling a story based on a diary written by a head nurse's working in a hospital during the pandemic. In her dairy, the head nurse documented fellow Wuhan medical professionals faithfully performing their duty while making personal sacrifice. These artistic works further boosted Chinese national pride because the success of China was contrasted with by public comparisons with other developed countries that had not effectively dealt with COVID-19 at the time.

3. Establishing 'Moral' Commitments: 'A Responsible Great Power'[62]

Along with presenting symbolic policy performance in both material and non-material forms, the Chinese government also utilised foreign policy to demonstrate its policy 'achievement' in assertively defending national sovereignty and acting 'responsibly' towards foreign affairs during the pandemic. This more assertive foreign policy exhibited during the pandemic has been justified as a 'moral commitment', combining the discourse of China's sovereign rights, 'victimhood', and global responsibility. While clearly aimed at China's neighbours and competitors, the policy also was directed toward diverting domestic attention away from public health issues by emphasizing the 'enemies' beyond the borders and by the promotion China's preferred values across the international community.

a. Assertive Foreign Policy: China is Being 'Wronged'

Recently China has justified an assertive foreign policy based on the continuing notion of 'victimhood' and as justified response to the threatening reactions of foreign powers who oppose its 'pacific' foreign policy. Chinese political discourse presented China as a 'victim' of foreign containment efforts while it is in fact merely excising its sovereign rights in a peaceful manner. This justification has the roots in the 'Century of Humiliation' national narrative and atrocities suffered in WWII. It is fuelled by the Japan's 'inadequate' apology for the war crimes as well as the continued American presence in the Asia Pacific region, especially American involvement in Taiwan.

This paradoxical policy setting, simultaneously assertive and emphasising victimhood was particularly reflected in China's foreign policy during the pandemic. For example, China was criticised of taking advantage of COVID-19 to consolidate its power and control in the South China Sea. In April 2020 a Chinese coast guard vessel sank a Vietnamese fishing boat in disputed waters; and in May the US Navy sent patrol ships in response to a Chinese survey and Coast Guard ships manoeuvring in the Malaysian Exclusive Economic Zone. Additionally, the Chinese aircraft carrier Liaoning conducted sea trials in disputed seas near the east coast of Taiwan; while China's newest aircraft carrier *Shandong* conducted her 'maiden sea trial' in May. In addition to these military actions, in April 2020 China's State Council decided to

---

[61] Xinhua, 'Jilupian tongxinzhanyi jiangbo' ['Documentary "Fighting Together" will be aired'], *Xinhuanet*, 31 August 2020, http://www.xinhuanet.com/politics/2020-08/31/c_1126435642.htm?baike (accessed on 4 December 2020).

[62] People.cn, 'Zhongguo kangyi zhangxian fuzeren daguo dandang', ['China demonstrates itself a responsible great power while combating the pandemic'], 19 March 2020, theory.people.com.cn/n1/2020/0319/c40531-31638571.html (accessed 28 August 2020).

establish two new districts in Sansha City, a prefecture level city which 'governs' the disputed territory in the South China Sea. While clearly a symbolic gesture, these geopolitical actions exhibited an intention to tighten control over the area to advance territorial claims, while adding additional impediments to Western efforts in support their preferred policy of freedom of navigation in the region.[63] During the pandemic, China's assertive postures in the South China Sea, (widely publicised in domestic media) worsened relations with Southeast Asian states, as well as the United States. US Secretary of State Mike Pompeo, echoed by Philippines' Foreign Affairs Secretary Teodoro Locsin, rejected China's 'historical rights' in the region and called for China to respect of 2016 Permanent Court of Arbitration's ruling which dismissed China's claims in the area.[64]

Reacting to this international criticism, the government reiterated its sovereign rights and emphasized its 'victimhood' in the face of 'anti-China' foreign forces. First, Chinese Minister of Foreign Affairs, Wang Yi, denounced the accusations that China was expanding in the South China Sea during the pandemic. He called such accusations 'preposterous' as 'China was fully focusing on cooperating with ASEAN states to combat the virus.' He noted that because of their joint efforts the ASEAN states and China were able to 'gain more political trust'. Wang further pointed out that it was 'shameful' and 'regrettable' that 'some countries' intended to 'sabotage' China's relations with ASEAN states, and 'endanger the peace and stability' in the region.[65] Second, in response to US and the Philippines' reference to the 2016 International Arbitration ruling, Chinese government reiterated that the 2016 decision was 'illegal, invalid and unacceptable;' as such US naval activities in the South China Sea were a 'violation of China's sovereignty and security' and a misuse of the *Convention on Law of the Sea*.[66] Given these stated 'sovereign rights', the establishment of new districts in Sansha City (which had led to international criticism), was 'reasonable, legitimate, and appropriate. Wang noted that those neighbouring states (Vietnam and the Philippines) which had criticised the extension of jurisdiction, had also previously built administrative structures and were 'in fact' engaged in acts of 'illegally seizing Chinese territory'.[67]

The application of 'victimhood' discourse has gathered much nationalist support: many netizens expressed appreciation of government's 'strong stand' internationally to defend China's national interests and handle 'US pressure'. This has

---

[63] Zachary Williams, 'China's tightening grasp in the South China Sea: A first-hand look', *The Diplomat*, 10 June 2020, https://thediplomat.com/2020/06/chinas-tightening-grasp-in-the-south-china-sea-a-first-hand-look/ (accessed 15 September 2020).

[64] Rahul Mishra, 'China's Self-Inflicted Wounds in the South China Sea', *The Diplomat*, 21 July 2020, https://thediplomat.com/2020/07/chinas-self-inflicted-wounds-in-the-south-china-sea// (accessed 15 September 2020).

[65] People.cn. 'Wang Yi: "Zhongguo liyong yiqing zai nanhai kuoda cunzai" shi wuji zhitan', ['Wang Yi: it is preposterous to say that China is taking advantages of the pandemic to expand the existence in South China Sea'], 24 May 2020, world.people.com.cn/n1/2020/0524/c1002-31721425.html (accessed 15 September 2020).

[66] Yu Xiaoqing and Yu Xiaoxuan, 'Yiqing zhixia meifang paichu junjian feiji pinfan zai nanhai zishi, waijiaobu duncu tingzhi' ['Ministry of Foreign Affairs requests US to stop sending warships and fight jets to stir South China Sea during the pandemic'], *Thepaper.cn*, 7 April 2020, https://m.thepaper.cn/newsDetail_forward_6860327 (accessed 15 September 2020).

[67] Liu Yanhua, 'Sansha shequ, heli hefa zheng dangshi' ['Establishing districts in Sansha is reasonable, legitimate, and good timing'], National Institute for South China Sea Studies, 21 May 2020, www.nanhai.org.cn/review_c/434.html (accessed 15 September 2020).

been a longstanding aspect of Chinese foreign policy discourse. Interestingly however, this victimhood" rhetoric has been extended not only in its resistance to its asserted territorial claims and economic disputes with other states but also in its response to international criticism as the alleged COVID-19 source country.

Chinese state media not only highlighted China as 'victim' of the coronavirus but also as a target of Western 'political manoeuvre'. During the high point of the pandemic there were calls for a global inquiry into the origin of the coronavirus and China's handling of the initial outbreak in Wuhan. Australia was one of early proponents of an independent investigation. In May 2020 Australia offered a draft motion to World Health Assembly requesting an evaluation of responses to the pandemic, which was supported by 122 countries.[68] The paradoxical victim-aggression policy was evident in China's reaction to Australia. On one hand, Chinese government criticised Australia for holding the 'ideological bias' and playing 'political manoeuvre' against China, which inevitably 'interfered [with] international cooperation' during the pandemic.[69] As the United States was one of the main virus infection sources in Australia, Chinese news outlets suggested that Australia should target the investigation towards the United States -- instead of 'using China as a scapegoat to appease its domestic public'.[70] On the other hand, China enacted a set of aggressive punitive measures against Australian economic interests. In May 2020, Chinese government imposed punishing tariffs on Australian exports such as barley and beef. In addition, the Chinese Ministry of Foreign Affairs advised Chinese citizens not to visit Australia due to the increasing 'racial discrimination and violence against Chinese' from 'local community, news media and law enforcement'.[71] For the same reasons, the Ministry of Education advised Chinese students not to choose Australia for tertiary education, an action that had significant adverse impacts on Australian universities. Employees in many state-own institutes received administrative orders to not to visit Australia for business or pleasure.[72]

Third, 'colonial and imperial victimhood' as it relates to Hong Kong has also dominated China's political discourse during the pandemic, rallying much domestic support against the 'foreign interference into Chinese domestic affairs'. On 30 June 2020, a new national security law entered into force in Hong Kong. The law has been widely criticised in international community for its vague definition of 'national security' and the lack of accountability and transparency. Many Western states and international organisations expressed deep concern over the law. Chinese government

---

[68] Daniel Hurst, 'Australia hails global support for independent coronavirus investigation', *The Guardian*, 18 May 2020, https://www.theguardian.com/world/2020/may/18/australia-wins-international-support-for-independent-coronavirus-inquiry (accessed 15 September 2020).

[69] Yu Xiaoqing and Wang Lunyu, 'Aodaliya youshui dui zhongguo yiqing zaoqi zhankai diaocha' ['Australia is lobbying to investigate China's handling of the initial virus outbreak'], *Thepaper.cn*, 23 April 2020, https://m.thepaper.cn/newsDetail_forward_7102830 (accessed 14 September 2020).

[70] CCTV.com, 'Aodaliya yao 'diaocha xinguan bingdu yuantou?' ['Australia wants to 'investigate the origin of the coronavirus'?'] 14 May 2020, m.news.cctv.com/2020/05/14/ARTIwt3p86y0VoY3QAnNwlyX200514.shtml (accessed 15 September 2020.

[71] People.cn, 'Zhu aodaliya shiguan tixing: zhongguo gongmin jinqi jinshen qianwang aodaliya' ['Reminder from the Chinese embassy in Australia: Chinese citizens should be cautious of planning to visit Australia'], 13 July 2020, http://travel.people.com.cn/n1/2020/0713/c41570-31781655.html (accessed 16 September 2020).

[72] Based on the author's personal communication with people working in a state-owned institute in Jiangsu Province, China, on 24 May 2020.

assertively responded to these criticisms citing its rights of sovereignty, security, and national development under international law. It stated that the security law aimed to 'protect Hong Kong residents' from 'separatist, terrorists, and foreign forces', and that the United States [and other states] should not interfere with China's domestic legislative action.[73] It has also asserted that China had performed all its obligations under the 'Sino-British Joint Declaration' that established the basis for Hong Kong return to China, noting that the Joint Declaration did not govern Hong Kong in 2020.[74]

The assertive foreign policy and policy discourse during the pandemic skilfully diverted Chinese public attention from domestic disease control and the increasingly onerous lockdowns to 'various threats overseas', including previous colonial powers, 'foreign forces' seeking to contain China, and 'aggressive' neighbour states. This foreign policy proffered to the domestic public that the 'true enemies' were the 'foreign forces out there' and by implication the coronavirus, despite its devastation was not a significant issue. For the international audience, however, the Chinese emphasis on the world's 'common enemy and shared victimhood' under COVID-19, was an attempt to mislead the international community from a more assertive foreign behaviour that sought to deepen Chinese foreign policy objective at a time when the world's attention was distracted by the pandemic.

b. 'Normative' Foreign Policy: China is Being A 'Responsible Power'

Chinese foreign was criticised during the pandemic for its failure to meet its responsibility.' This responsibility related to it its lack of transparency, accountability, and protection of individual rights during the crisis. China countered these criticisms by emphasizing its state responsibility through it implantation of a 'responsible public policy' of reporting and controlling the disease. And more importantly with its widely publicised provision of global common goods during the pandemic. To address the criticism on lack of transparency, the government claimed that China's public health institutes have always been 'open, transparent, and responsible' in terms of information sharing, having reported then yet 'unknown virus' to WHO on 3 January 2020. In support of its position of as a responsible power, Chinese state media noted China has made a 'great contribution' to world public health in combatting COVID-19 as it has 'efficiently contained the virus spread' and made 'tremendous economic sacrifice' with a nation-wide lockdown.[75] It argue that in contrast other countries such the United States, have acted neither appropriately nor responsibly. From the Chinese perspective, these countries not only did not slow transmission rates but also violated their

[73] People.cn, 'Waijiaobu: jiang dui waibuy shili ganshe xianggang xingjing yuyi fanzhi' ['Ministry of Foreign Affairs: China will fight against any foreign interference in Hong Kong affairs'], 27 May 2020, world.people.com.cn/n1/2020/0527/c1002-31726340.html (accessed 14 September 2020).
[74] People.cn, 'Waijiaobu bo meifang ganshe zhongguo shegang lifang' ['Ministry of Foreign Affairs criticised US interference in Hong Kong's law-making'], 25 May 2020, world.people.com.cn/n1/2020/0525/c1002-31723142.html 1/ (accessed 14 September 2020).
[75] People.cn, 'Quanmin zhan "yi", zhongguo dui shijie de dandang' ['People's war against coronavirus: China's responsibility for the world'], 9 February 2020, http://opinion.people.com.cn/gb/n1/2020/0209/c223228-31578105.html (accessed 17 September 2020).

international responsibilities, spreading what China has called a 'political virus' by 'sabotaging other countries' genuine efforts' to combat the disease.[76]

Furthermore, Chinese government attempted to burnish its much advertised 'responsibility power' image by providing global commons goods during the pandemic. First, according to the 2020 White Paper, China had offered a large amounts of humanitarian aid including USD$50 million cash to WHO, sending medical teams to 27 countries, and delivering medical aid to 150 countries and 4 international organisations.[77] For example, on 23 March 2020 a Chinese chartered plane arrived in Italy and delivered 155 ECMOs, 1.1 million FFP2 and N95 face masks, 305,000 surgical masks, 205,000 gloves, 1,000 COVID-19 test kits, and Personal Protection Equipment, a lot of which were donation from Chinese government and Chinese businesses.[78] Other less developed and neighbouring countries also received donations or medical aid from China.[79] Chinese media contrasted its 'altruism' with Taiwan's 'selfishness' as Taiwan refused to allow face mask exports to the PRC, when it donated 100,000 N95 face masks to Australia in January 2020 for use in the widespread bushfires.[80]

Second, Chinese government announced its willingness to cooperate and share the vaccine it had developed with the less developed countries. Two Chinese vaccine makers conducted trials with Pakistan National Institute of Health where people were reported to be eager to receive the vaccines.[81] As President Xi stated at World Health Assembly in May 2021, the domestically-made Chinese vaccine was envisioned to become a 'global public good'. And China would ensure its 'accessibility and affordability in developing countries'.[82] This global outreach was not without effect. China's vaccination research and its stated intention to share the results have eased diplomatic ties with its Southeast Asian neighbours. Standing in contrast to the 'America First' policy under the Trump Administration, the policy signalled China's continued aspirations for global leadership in the post-pandemic world order.

Third, Chinese government stated its intention to participate in world economic recovery and shape the post-pandemic world system with its power, influence, and preferred values. The 'One Belt One Road' Initiative, China's global infrastructure development strategy remained the centrepiece of Chinese foreign policy related to

---

[76] People.cn, 'Zhongguo daibiao zai lianda quanhui yanli bochi meifang wuduan zhize' ['Chinese UN representatives denounced US criticism at General Assembly'], 12 September 2020, http://world.people.com.cn/gb/n1/2020/0912/c1002-31858953.html (accessed 17 September 2020).

[77] China's State Council Information Office, 'Fighting COVID-19: China in action'.

[78] People.cn, 'Dapi wuzhi zi zhongguo dida milan, yuanzhu yidali kangji yiqing' ['Chinese resources arrived in Milan to assist Italy in the pandemic'], 24 March 2020, http://world.people.com.cn/n1/2020/0325/c1002-31647279.html (accessed 18 September 2020).

[79] People.cn, 'Zhongguo yuanzhu duoguo kangyi' ['China gives aids to multiple countries to combat the pandemic'], 6 April 2020, http://paper.people.com.cn/rmrbhwb/html/2020-04/06/content_1980222.htm (accessed 18 September 2020).

[80] Wang Ping, 'Jie yiqing la chouhen tai dangju shihe juxin' ['Using the pandemic to flaming hatred: what is Taiwan's intention?'], *People.cn*, 13 February 2020, http://tw.people.com.cn/n1/2020/0213/c14657-31584337.html (accessed 18 September 2020).

[81] Sui-Lee Wee, 'From Asia to African, China promotes its vaccines to win friends', *New York Times*, 11 September 2020, https://www.nytimes.com/2020/09/11/business/china-vaccine-diplomacy.html (accessed 18 September 2020).

[82] Xinhuanet.com, 'China's COVID-19 vaccine to become global public good when available: Xi', 18 May 2020, http://www.xinhuanet.com/english/2020-05/18/c_139066851.html (accessed 18 September 2020).

overseas investment and economic interactions. As President Xi's main policy 'innovation' and achievement, the Initiative has often been used as a basis for economic cooperation after the pandemic, despite evidence that it may not have sufficient funding.[83] Nevertheless, President Xi described the Initiative 'the answer' to the many challenges in post-pandemic world where China 'works with its partners' to build 'a road to multilateral cooperation, public health, economic revival, and full development potential.'[84]

Besides the One Belt One Road Initiative, China has also pushed its own preferred values in a 'global community of shared future'. This concept echoes the idea of 'Beijing Consensus', a Chinese developmental model featuring 'stable yet repressive politics and high-speed economic growth'. Unlike 'Beijing Consensus' that was endogenously defined and never fully embraced in Chinese official discourse, 'a global community of shared future' often appeared in foreign policy rhetoric during the pandemic. Though not particularly clear, the concept involves two layers of meaning. First, China's political system has 'advantages', which were 'evident' in the pandemic and assisted in China fulfilling its 'international obligations.' Therefore, foreign powers should refrain from intervening in its affairs. [85] This non-intervention principle includes the idea that normative values and human rights can be interpreted and implemented differently by different political systems. For example, Chinese delegates insisted that 'people's happy life was their primary human right' at the UN Human Rights Council in September 2020.[86] Second, whether or not 'participating [in] multilateralism and global cooperation',[87] China intends to play a more important role in global issues in the post-pandemic world.

CONCLUSION

This paper explores China's public policy during the COVID-19 pandemic. Drawing upon the scholarship of performance-based policy that focuses on policy effectiveness, regime legitimacy, and state capacity building, this article argues that the Chinese government and Chinese Communist Party maintained social stability and domestic support during the outbreak by exercising a performative public policy. This policy approach emphasized the construction and positive presentation of policy achievement, in both material and non-material forms. From this perspective, symbolic performance can be as important as a concrete policy -- and in certain circumstances more effective, such as during the early stages of COVID-19 where a paucity of scientific knowledge created difficulties in evaluating the effectiveness of a particular

[83] Plamen Tonchev, 'The Belt and Road after COVID-19', *The Diplomat*, 7 April 2020, https://thediplomat.com/2020/04/the-belt-and-road-after-covid-19/ (accessed 21 September 2020).
[84] China.com, 'Jujiao hou yiqing shidai, xi jinping wei zhe tiao "lu" fuyu xin neihan' ['In post-pandemic era, Xi Jinping gave the 'Road' new meaning'], 20 June 2020, http://news.china.com.cn/2020-06/20/content_76183791.htm (accessed 21 September 2020).
[85] People.cn. 'Zhongguo zhu fayu dashi: yidai yilu shi guoji huhui hezuo de changyuan jihua' ['Chinese Ambassador in France: One Belt One Road is a long-term international cooperation initiative'], 27 August 2020, world.people.com.cn/n1/2020/0827/c1002-31838589.html (accessed 21 September 2020).
[86] People.cn, 'Zhongguo daibiao: renmin de xingfu shenghuo jiushi zuida de renquan' ['Chinese delegates: People's happy life is the primary human rights'], 19 September 2020, world.people.com.cn/n1/2020/0919/c1002-31867495.html (accessed 21 September 2020).
[87] Li Jiabao, 'Zhongguo kangyi chengguo xiang shijie chuandi xinxi' ['China's achievement of combatting the pandemic conveys confidence to the world'], *People.cn*, 14 September 2020, world.people.com.cn/n1/2020/0914/c1002-31859878.html (accessed 21 September 2020).

health initiative. While this piece has focused on China, performance public policy is found in other countries' policymaking as well. For example, US President Trump's in 2020 who took performance and political theatre to new heights in his tenure, commented on American economy rebounding and COVID-19 receding; while at the same time the US death toll climbed and there were a record number of unemployed. Similarly, in Australia, the national border was closed to Chinese travellers early in the pandemic (February 2020) alleged to reassure the domestic public; nevertheless, the main infection sources in Australia were from Europe and America.

China's performance public policy during the pandemic had three elements. First, the government constructed and presented policy outcomes, regardless of their actual success or failure, in a positive light. This positive policy achievement was largely credited to individual efforts President Xi. Second, the government provided political goods, such as national pride, law and order, and shared Chinese values, to optimize the positive public perception of government performance. The political goods were often accompanied by the anti-Western and nationalist discourse. Third, Chinese government attempted to promote its 'moral commitments' in foreign policy. Adopting a 'victimhood' rhetoric, the government deployed an assertive foreign policy to 'protect' security and sovereignty against the 'anti-China foreign forces' during the pandemic. The assertive policy diverted domestic attention from a spreading disease to 'threats overseas', and rallied nationalist support in the issues such as South China Sea and Hong Kong. Furthermore, Chinese government has also announced its intention to assume a more important role in a post-pandemic world through the provision of global public goods, such as sharing vaccinations and stimulating recovery. China has also aimed to shape the international normative community with its own preferred values embedded in an authoritarian political culture.

STATE-BASED ONLINE RESTRICTIONS: AGE-VERIFICATION AND THE VPN OBSTACLE IN THE LAW

Youssef A. Kishk*

Abstract: Since the inception of the internet, the availability of online pornography to minors has been a major concern, and the federal government has tried and failed to effectively prohibit minors' access to these materials online. Some states have enacted legislation to force commercial entities distributing this harmful material online to enact reasonable age-verification. These "porn" statutes may be subject to constitutional challenges on the basis of overbreadth and privacy. Outside of potential constitutional challenges, these laws are indicative of a potential national trend in state-created online pornography restrictions, and the issues of ineffectiveness and inconsistency present within the laws themselves merit an analysis. Additionally, this paper will use these recent laws and their issues as a basis to explain the place of virtual private networks ("VPNs") in the law. Particularly, VPNs are the most common method to circumvent state-enforced online regulation and yet they tend to be ignored or overlooked by statutes despite their popularity. Causing VPN companies to profit from these "porn" law restrictions, by giving online, and potentially minor, users the ability to ignore most age-verification measures put in place by these laws.

Keywords: State-Based; Online Restrictions; Age-Verification; VPN

* University of Mississippi School of Law, United States.

Table of Contents

INTRODUCTION

There has been a recent wave of state-enacted laws mandating the incorporation of more stringent age-verification systems on websites that distribute material harmful to minors online.[1] These laws are essentially modern state attempts of the Child Online Protection Act, that attempt to detect where an online user is searching from to enforce the relevant state's required age-verification.[2] While these "online pornography restriction" laws have been passed in multiple states, such as Louisiana, Virginia, Arkansas, and Utah,[3] for the purposes of this paper, Mississippi's law will be the primary example evaluated.[4] But the critiques and concerns about the Mississippi law will be applicable to most of the other equivalent state statutes.

On top of constitutional concerns regarding these laws, they are a characteristic example of how the American legal system tends to treat virtual private networks, which is to ignore them. These laws show how despite the impact virtual private networks can have on the enforcement of these and other laws, virtual private networks are routinely ignored, or allowed to skirt through vague legal provisions, that may or may not apply to them.[5] This ignorance towards VPNs and their potential uses must cease to increase the effectiveness of laws in an online context, while addressing concerns of users online who may be unsure on the legality of specific VPN uses.

The Mississippi law expressly prohibits internet service providers, and search engines from liability under this law, so long as these excluded entities are not directly responsible for the creation of "material harmful to minors."[6] The following discussion will not cover the excluded entities, even if their involvement may affect the created regulations.

Critics of these laws argue that they violate the individual's right to privacy.[7] But while aware of the possible privacy issues surrounding this law, this paper will not

---

[1] *See generally* Miss. Code Ann. § 11-77-5; A.C.A. § 4-88-1305*;* La. R.S. § 51:2121.

[2] *See generally* ERIC N. HOLMES, CONG. RSCH. SERV., R47049, CHILDREN AND THE INTERNET: LEGAL CONSIDERATIONS IN RESTRICTING ACCESS TO CONTENT 9-10 (2022) (outlining the Child Online Protection Act and why it was passed).

[3] *See* Marc Novicoff, *A Simple Law Is Doing the Impossible. It's Making the Online Porn Industry Retreat.*, POLITICO (Aug. 8, 2023, 4:30 AM), https://www.politico.com/news/magazine/2023/08/08/age-law-online-porn-00110148.

[4] *See* Miss. Code Ann. §§ 11-77-1 – 11-77-7.

[5] *See* Kyle Berry*, This Content is Unavailable in Your Geographic Region: The United States' and the European Union's Implementation of Anti-Circumvention Measures*, 55 VAND. J. TRANSNAT'L. 485, 517 (2022) (there are circuit splits in the United States on if the act of circumvention, like when using VPNs to change your online location, "is sufficient for liability or whether the act of circumvention must be connected to an act of infringement."). This ambiguity allows VPNs to facilitate infringing acts without legal repercussions because they are not adequately addressed by the American legal system, even when their use directly inhibits a statute's goal such as in the case of the Mississippi statute. *See* Miss. Code Ann. §§ 11-77-1 – 11-77-7 (VPNs are not addressed in the statute).

[6] Miss. Code Ann. § 11-77-7.

[7] Lacey Alexander, *Pornhub blocks access in Mississippi in response to new law*, MISS. PUB. BROAD. (Jul. 5, 2023), https://www.mpbonline.org/blogs/news/pornhub-blocks-access-in-mississippi-in-response-to-new-law/.

delve deeply into the matter. Especially as lawsuits on the issue have already been filed, this paper will not address them in depth.[8]

The Mississippi law will be analyzed in the following ways. Part II will outline the elements of the law,[9] the technology required to enforce the law, the effects of the contemplated and not contemplated technology used, and the common responses to the law's enactment. Part III will consist of constitutional challenges that the law will likely be subject to, including the constitutionality of banning the narrower category of obscenity online, a first amendment facial challenge, a constitutional argument for the law's underinclusivity and overinclusivity in achieving its compelling government interest. And Part IV will address the broad concerns of these laws through the lens of a national trend, issues with specific provisions of the Mississippi law, and the technological concerns tied to the law, including what it highlights about the use of virtual private networks and how they are generally treated in the law.

I. BACKGROUND

A. Mississippi's Online Age-Verification Law Explained

Senate Bill No. 2346, now classified as Miss. Code Ann. §§ 11-77-1 – 11-77-7 (subsequently referred to as the "Mississippi law" or "Mississippi statute" in this paper), took effect in July, 2023, making commercial entities who distribute "material harmful to minors" online liable to the individual for damages resulting from a minor accessing their website, if the commercial entity fails to perform "reasonable age-verification."[10] This age-verification is to prevent minors from accessing these platforms online, so long as the website is made up of a "substantial portion" of this "material harmful to minors."[11] Commercial entities or third parties performing this "reasonable age-verification" are not to keep any identifying information collected, once a user's age is verified, and is granted access to the restricted website.[12]

Affected commercial entities that do not comply can be liable to an individual for damages a minor sustained from accessing their platform, this may include court costs and attorney fees.[13] Minors are any individual under the age of eighteen, despite the age of consent in Mississippi being sixteen years of age.[14] The statute's goal is to restrict the access of minors, not adults, to harmful material online.[15]

1. "Material Harmful to Minors"

The statute restricts more than just obscenity.[16] The statute's definition of "material harmful to minors" uses very similar language as the three-pronged *Miller*

---

[8] *See e.g.*, Free Speech Coal., Inc. v. Anderson, 2023 U.S. Dist. LEXIS 134645 (D. Utah Aug. 1, 2023).
[9] Particularly that while advertised as a porn restriction, the statute restricts the much broader category of "material harmful to minors." *See* Miss. Code Ann. §§ 11-77-3 – 11-77-5.
[10] Miss. Code Ann. § 11-77-5.
[11] *Id.*
[12] *Id.*
[13] *Id.*
[14] *Id.* at § 11-77-3; Miss. Code Ann. § 97-3-65 (there is no statutory rape charge in Mississippi if the younger individual is sixteen years of age or older, indicating sixteen is the state's age of consent).
[15] *See* Miss. Code Ann. § 11-77-5.
[16] *See Id.* at § 11-77-3.

obscenity test, but broadens its scope to relate to the sensibilities of minors.[17] Even content that has serious value and is excluded from this harmful material definition is framed in how it applies to minors, and not all individuals.[18] This harmful material definition coincides with the doctrine from *Ginsberg v. State of N.Y.*, that non-obscene material for adults, can be regulated for minors if it is considered harmful to them, also known as "variable obscenity."[19] This is broader than obscenity for adults, but as it concerns children, the government has greater power to restrict content that falls under this variable obscenity scope.[20] The Mississippi statute's definition of "material harmful to minors" is essentially applying *Ginsberg* "variable obscenity" in a modern online setting, where it faces difficulties that were not present when *Ginsberg* was decided.

Additionally, the serious value exception to "material harmful to minors" is difficult to apply in an online context depending on the material at hand.[21] The exact meaning and scope of the serious value exception has not been sufficiently determined, especially when restricting content for minors.[22] This vagueness is compounded when it comes to evaluating new technology, and whether its use have serious value or not, such as the use of deepfakes to create pornographic content, which some argue inherently has serious technological value, but the existence of this debate shows how problematic applying the serious value exception can become.[23] The Mississippi statute currently leaves this exception quite vague. Which can make it difficult for those who want to seek damages against a commercial entity, as they may be unsure if content qualifies, or for commercial entities who may not know if the content they distribute falls under this serious value exception.

---

[17] *Compare* Miller v. California, 413 U.S. 15, 24 (1973) (the *Miller* test is: "(a) whether 'the average person, applying contemporary community standards' would find that the work, taken as a whole, appeals to the prurient interest …; (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value."), *with* Miss. Code Ann. § 11-77-3 (the Mississippi law defines material harmful to minors as: "(i) Any material that the average person, applying contemporary community standards would find, taking the material as a whole and with respect to minors, is designed to appeal to, or is designed to pander to, the prurient interest; (ii) Any of the following material that exploits, is devoted to, or principally consists of descriptions of actual, simulated, or animated display or depiction of any of the following, in a manner patently offensive with respect to minors…; and (iii) The material taken as a whole lacks serious literary, artistic, political, or scientific value for minors.").

[18] *Compare* Miller v. California, 413 U.S. 15, 24 (1973) (The *Miller* test's exception to obscenity is "whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value."), *with* Miss. Code Ann. § 11-77-3 (The Mississippi law's exception to material harmful to minors is "[t]he material taken as a whole lacks serious literary, artistic, political, or scientific value for **minors**.") (emphasis added).

[19] *See* Ginsberg v. State of N. Y., 390 U.S. 629, 631-34, 636, 673 (1968) (The court upheld a verdict that a store owner was guilty of violating a New York penal statute for selling a 16-year-old boy a magazine that was obscene for minors, but not obscene for adults.).

[20] *Id.* at 636.

[21] *See* Miss. Code Ann. § 11-77-3.

[22] *See* Todd E. Pettys, *Serious Value, Prurient Appeal, and "Obscene" Books in the Hands of Children*, 31 WM. & MARY BILL OF RTS. J. 1003, 1040 (2023).

[23] *See* Bradley Waldstreicher, *Deeply Fake, Deeply Disturbing, Deeply Constitutional: Why the First Amendment Likely Protects the Creation of Pornographic Deepfakes*, 42 CARDOZO. L. REV. 729, 755-57 (2021).

2. Individual Damages

The Mississippi law holds commercial entities liable for the individual damages that a minor can accrue from consuming "material harmful to minors" commercial entities distribute without employing age-verification.[24] The language of the law indicates this liability is to the individual harmed or their representative, and to seek damages from a specific commercial entity, then that individual must pursue damages in court.[25] The law creates a cause of action against these commercial entities.[26]

While the law treats this as a private cause of action, there is a possibility the state can pursue action itself.[27] The "parens patriae" doctrine may be an avenue for the State to sue non-complying commercial entities on behalf of their citizens.[28] As the Mississippi statute highlights the state's compelling interest to protect minors from accessing restricted material, and may give the state third-party standing to sue affected commercial entities.[29]

3. "Reasonable" Age-Verification

The Mississippi law is not a complete bar to the distribution of "material harmful to minors" online, but it restricts the access of minors to this content by mandating the use of "reasonable age-verification."[30] This verifies whether a user is a minor, and thus barred, or an adult and allowed access to an affected website. Commercial entities verify user ages by collecting their personal information, to verify the user's true age.[31] Commercial entities are liable if they are found retaining any of this identifying information, or if they fail to use reasonable age-verification, which should be more thorough than the "honor-system" frequently used today.[32]

B. Age-Verification Technology

Reasonable age-verification needs users to provide proof of their age.[33] This requires two technical components to function: the use of geoblocking and providing personal information to an online party.[34]

---

[24] Miss. Code Ann. § 11-77-5.
[25] *See Id.*
[26] *See Id.*
[27] *See* Alexander, *supra* note 7 (reporting commercial entities may face fines from the attorney general for not complying with the Mississippi statute).
[28] See Seth Davis*, Implied Public Rights of Action*, 114 COLUM. L. REV. 1, 44 (2014).
[29] *See Id.* at 22-23.
[30] Miss. Code Ann. § 11-77-5 ("'Reasonable age verification methods' include verifying that the person seeking to access the material is eighteen (18) years of age or older by using any of the following methods: (i) Provide a digitized identification card; (ii) Require the person attempting to access the material to comply with a commercial age verification system that verifies in … the following ways: 1. Government-issued identification; or 2. Any commercially reasonable method … to verify the age of the person….").
[31] *Id.* at § 11-77-3.
[32] See *Id.* at § 11-77-5; Christine Marsden, *Age-Verification Laws in the Era of Digital Privacy*, 10 NAT'L. SEC. L.J. 210, 214 (2023).
[33] Miss. Code Ann. § 11-77-3.
[34] *See* Tal Kra-Oz, *Geoblocking and the Legality of Circumvention*, 57 IDEA 385, 388 (2017); Byrin Romney*, Screens, Teens, and Porn Scenes: Legislative Approaches to Protecting Youth from Exposure to Pornography*, 45 VT. L. REV. 43, 68-69 (2020).

1. Geoblocking Explained

Geoblocking is technology used to locate the approximate geographic location of an online user,[35] and to restrict their access to certain content based on their physical location.[36] The geoblocking component of age-verification is the most common method commercial entities use to assess if a user is coming from the state of Mississippi to then apply the Mississippi law's age-verification.[37] As not all places in the country or the world require these entities to enforce stringent age-verification, so the commercial entities use geoblocking to identify the approximate location of users through the user's IP address.[38] If the IP address is found to come from Mississippi, then users who access the website are redirected to the age-verification system, or equivalent response, established by the website.[39] This redirection only occurs to users identified as being physically located in states mandating age verification, to prevent one state's law from affecting out-of-state users of the website.[40]

2. Personal Information Requirement

The other component of age-verification, is that once redirected, the commercial entity must verify the user's age by evaluating the user's personal information. The user provides a form of valid identification to the age-verification system, which requires the user to have some government or digital ID present in the database's system.[41] Once user age is determined, minors are barred access, but adult users can access restricted websites.[42]

Privacy concerns were contemplated, so the Mississippi statute makes commercial entities liable if they retain any identifying information of a user after the age-verification.[43] Some commercial entities have banned all Mississippi users from accessing their website due to privacy concerns.[44] These privacy concerns regarding personal identifying information persist, despite the use of identification to verify age required to access other material in society such as "purchasing alcohol."[45]

C. Age-Verification Circumvention

1. Virtual Private Networks

A Virtual Private Network (VPN) is a service internet users can employ to increase the privacy and the protection of their online activities, especially when using

---

[35] *See* Marketa Trimble, *The Future of Cybertravel: Legal Implications of the Evasion of Geolocation*, 22 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 567, 585-86 (2012) ("Trimble I").
[36] Peter K. Yu, *A Hater's Guide to Geoblocking*, 25 B.U.J. SCI. & TECH. L. 503, 504 (2019).
[37] *See* Marketa Trimble, *Copyright and Geoblocking: The Consequences of Eliminating Geoblocking*, 25 B.U.J. SCI. & TECH. L. 476, 483 (2019) ("Trimble II").
[38] *See* Alexander, *supra* note 7.
[39] *Id.*
[40] *See Id.*
[41] Miss. Code Ann. § 11-77-3.
[42] *See Id.* at § 11-77-5.
[43] *Id.*
[44] *See* Alexander, *supra* note 7 ("'We are sorry to let our loyal visitors in these states down but we have opted to comply with the newly effective law in this way because it is ineffective and worse, will put both user privacy and children at risk.' Pornhub said in a tweet.").
[45] Marsden, *supra* note 32 at 239.

public Wi-Fi.[46] These VPNs allow their users to hide their IP address location, to circumvent geoblocks that companies, such as Netflix, and commercial entities affected by the Mississippi statute employ.[47] Meaning a VPN allows an internet user, located in Mississippi, to appear as if they are in a different state or nation, allowing them to access content restricted in Mississippi but not in their pretend location.[48] Using VPNs to circumvent geoblocks put into place by commercial entities would allow Mississippi users to disregard any age-verification mandated by the Mississippi statute, unless the website applies age-verification to all of its users regardless of location.[49] Following the passage of Virginia's age-verification law, and "after Pornhub pulled out of Virginia, searches for VPNs spiked in the state," suggesting a connection between the use of geoblocking and increases in the use of VPNs.[50]

2. Effects of Circumvention

When users alter their online location with a VPN, the user is subject to geoblocks and content specified for the location their IP address is disguised as coming from.[51] VPN services heavily advertise these "cybertraveling" features.[52] Most popular VPN services can be used for under three dollars a month, and less secure free VPNs can be used and obtained without any age-verification.[53] Under the Mississippi law, user circumvention of age-verification systems and VPNs are not addressed, so a commercial entity is still in compliance if they instituted age-verification for users flagged as coming from Mississippi.[54] Leading to situations, where a minor is harmed from accessing "material harmful to minors" distributed online by using a VPN, but so long as the commercial entity is complying with the Mississippi statute, then the minor has no culpable party to obtain damages for their injuries.

D. Commercial Entity Response

In response to the Mississippi law, affected commercial entities responded in a few ways. The first response was to use geoblocking systems, and to enforce age-verification on all users found coming from Mississippi.[55] While this likely occurred

---

[46] Steve Symanovich, *What is a VPN?*, NORTON (Feb. 24, 2022), https://us.norton.com/internetsecurity-privacy-what-is-a-vpn.html ("A virtual private network... gives you online privacy and anonymity by creating a private network from a public internet connection. VPNs mask your internet protocol (IP) address so your online actions are virtually untraceable. … VPN services establish secure and encrypted connections to provide greater privacy …. A VPN creates a type of tunnel that hides your online activity… so that cybercriminals, businesses, government agencies, or other snoops can't see it.").

[47] Berry, *supra* note 5 at 488-89.

[48] *See* Sabrina Earle, *The Battle against Geo-Blocking: The Consumer Strikes Back,* 15 RICH. J. GLOBAL L. & BUS. 1, 11 (2016).

[49] *See* Romney, *supra* note 34 at 72.

[50] *See* Novicoff, *supra* note 3.

[51] *See* Earle, *supra* note 48 at 11.

[52] *See generally* Trimble I, *supra* note 35.

[53] Best VPNs of September 2023, USA TODAY (June 21, 2023, 9:55 AM), https://www.usatoday.com/money/blueprint/l/best-vpns/?utm_content=150585994043&utm_term=kwd-320157419234&utm_campaign=20344709965&gclid=Cj0KCQjwl8anBhCFARIsAKbbpyT4eTusULLebNGtLY-cwfZlVghg__n29ArFUgL3HzbBWWsRVMSsP7kaAgq7EALw_wcB; *See e.g.* Lawrence Wachira, *How to Unblock Porn Sites From Anywhere in 2023*, VPN MENTOR (Sept. 15, 2023), https://www.vpnmentor.com/blog/how-to-unblock-porn-sites-from-anywhere/.

[54] *See* Miss. Code Ann. § 11-77-5.

[55] *See Id.*

during the statute's enactment, this has fallen out of favor and has been mainly replaced by the next response.[56]

The second response was to completely block all Mississippi users from accessing affected websites. The commercial entities use geoblocking, but instead of verifying user age, they ban all Mississippi users from accessing their websites.[57] This is a form of malicious compliance to the Mississippi statute. This has become the common response for many of the larger commercial entities that are easier to hold liable.[58] Especially for the commercial entities that produce and distribute harmful material, such as the various affiliates of "MindGeek," the largest pornography company globally.[59]

The final response is non-compliance with Mississippi's statute, allowing users from anywhere to access their website with little to no age-verification.[60] Since larger companies complied with Mississippi's law, "Pornhub… claims that traffic soared for its noncompliant competitors."[61] While no data was provided for this claim, if true, then smaller commercial entities distributing "material harmful to minors" may risk lawsuits for their non-compliance in exchange for greater traffic and profit on their platforms. Unless the Mississippi law can make the cost of non-compliance severe enough to warrant enacting age-verification systems, then many affected commercial entities will continue not to comply for greater profits.

II. CONSTITUTIONAL CHALLENGES

The Mississippi law likely violates the first amendment's freedom of expression, incorporated under the fourteenth amendment.[62] This analysis will entail evaluating the constitutionality of completely banning obscenity online, a facial challenge to the law, and an analysis of the law's underinclusive and overinclusive restrictions.

While online activity is usually private action that receives stronger constitutional protection, the distribution of material harmful to minors by commercial entities online will likely be considered public rather than private. As the act of

---

[56] *See* Courtney Ann Jackson, Two new Mississippi laws are designed to protect kids from easy access to porn, WLBT 3 (Jul. 3, 2023, 9:32 PM), https://www.wlbt.com/2023/07/04/two-new-mississippi-laws-are-designed-protect-kids-easy-access-porn/.

[57] Alexander, *supra* note 7.

[58] *See Id.*

[59] *See* Joe Castaldo*, Lifting the veil of secrecy on MindGeek's online pornography empire*, THE GLOBE & MAIL (Feb. 4, 2021), https://www.theglobeandmail.com/business/article-mindgeeks-business-practices-under-srutiny-as-political-pressure/#:~:text=MindGeek%2C%20which%20operates%20from%20Montreal,revenue%20and%201%2C800%20employees%20globally. When Mississippi users access the top pornography companies' websites, they are directed to a video explaining users in the state are banned from accessing their website in response to the law. *See* Novicoff, *supra* note 3.

[60] *See* Meghan Mcintyre, *Many pornography websites aren't complying with new Va. age verification law*, VA. MERCURY (Aug. 23, 2023, 12:04 AM), https://www.virginiamercury.com/2023/08/23/many-pornography-websites-arent-complying-with-new-va-age-verification-law/. (Non-compliance reported with Virginia's law, also occurs for Mississippi's law.).

[61] Novicoff, *supra* note 3.

[62] *See* U.S. Const. amend. I; U.S. Const. amend. XIV, § 1.

purchasing or accessing this material online could be private to individual users.[63] But the distribution of content online goes beyond private possession for the commercial entities, and since the State has wide discretion in regulating obscenity, the distribution of content online could be regulated as public action.[64] A public action determination gives the State more leeway in restricting material distributed by these commercial entities regardless of enacted disclaimers or age-verification.[65] Public action can be regulated at greater levels than purely private action at home, even if not to the level as a "place of public accommodation."[66]

A. Possibility of Complete Online Obscenity Ban

Before addressing the Mississippi law's constitutionality, an analysis on restricting the broader concept of online obscenity is required. As if a complete restriction on online obscenity is unconstitutional, then restricting "material harmful to minors," which is broader than obscenity, would be unconstitutional.

Under *Miller*, anything found to be obscene is not protected by the first amendment, and the government can regulate as it sees fit.[67] Meaning the government can likely regulate purely online obscenity, if the restricted content is obscene under the *Miller* test.[68] As states are given wide discretion when regulating obscenity due to the lack of constitutional protection given to obscene speech.[69]

The exception to this government interest, would be if the online obscenity falls under the right to sexual privacy.[70] But even if the government still recognizes a right to sexual privacy in the context of online obscenity, "the government would remain free to enforce obscenity statutes for publicly distributed obscene" online material.[71] So the Mississippi law is constitutional when regulating obscene content, but not necessarily for all "material harmful to minors," as the Mississippi statute only restricts the distribution of this content not the possession of it.[72]

Alternatively, it is possible that while the State can restrict obscene material online, non-complying commercial entities could escape liability under Section 230 of the Communications Decency Act of 1996 (CDA).[73] Which states "[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[74] And since the Mississippi law relates to commercial entities distributing "material harmful to minors" online, the Mississippi law may violate this provision of the CDA, as it attempts to hold

---

[63] *See* Jennifer M. Kinsley, *Sexual Privacy in the Internet Age: How Substantive Due Process Protects Online Obscenity*, 16 VAND. J. ENT. & TECH. L. 103, 117 (2013).
[64] *See* Stanley v. Georgia, 394 U.S. 557, 568 (1969).
[65] *See* Paris Adult Theatre I v. Slaton, 413 U.S. 49, 59 (1973).
[66] *Id.* at 57 & 69.
[67] *See generally* Miller v. California, 413 U.S. 15 (1973).
[68] *See Id.* at 24.
[69] Roth v. United States, 354 U.S. 476, 481 (1957).
[70] *See e.g.*, Lawrence v. Texas, 539 U.S. 558 (2003).
[71] Kinsley, *supra* note 63 at 131.
[72] Miss. Code Ann. § 11-77-5.
[73] *See* 47 U.S.C. § 230.
[74] *Id.* at § 230(c)(1).

commercial entities liable for the distribution of content, if posted by third-party creators. If this occurs, the Mississippi law would likely be invalid under Federal law.

B. Facial Challenge

1. Public Right of Action

The Mississippi statute holds non-complying commercial entities that knowingly distribute "material harmful to minors" liable.[75] This civil liability can be pursued by the minors, or their representatives, who are damaged from exposure to restricted content on a non-complying website.[76] This may be construed as a private right of action, where the government allows litigation to occur without government action.[77]

This is really an implied public right of action, as while the State is not seeking damages, the statute's mere existence promotes the State's interest of restricting "material harmful to minors."[78] The Mississippi statute forces commercial entities to follow this government restriction or risk lawsuits. The statute could give the State standing to sue non-abiding commercial entities on behalf of damaged individuals, under the "parens patriae" doctrine.[79] Here, the State may argue that its compelling interests in protecting minors from harmful material online gives it third-party standing[80] to sue non-complying entities. Transforming a private cause of action into essentially a government fine through litigation. The statute does not mention government action in response to non-complying entities, but the lack of disclaimer on potential state action creates implications of potential government action, rather than private action. Overall, this means any arguments attempting to posit the Mississippi statute is not state action, are misleading and attempting to discount the State's involvement in restricting "material harmful to minors" online.

2. Overbreadth

The goal of the Mississippi law is to prevent minors from accessing harmful online content.[81] The law requires adults to provide commercial entities with identifying information to access "material harmful to minors," so adults are not prevented from accessing this content.[82] This may not be an issue if the law outright banned pornography in Mississippi, but since the law limits this restriction to minors, possible overbreadth by the law, and an impermissible burden on the rights of adults may be a concern.[83]

---

[75] Miss. Code Ann. § 11-77-5.
[76] *Id.*
[77] *See* Davis, *supra* note 28 at 71.
[78] *See Id.* at 9 & 17.
[79] *Id.* at 44 ("In Massachusetts v. Bull HN Information Systems, Inc., for example, the district court held that the Age Discrimination in Employment Act authorizes a parens patriae suit by defining a 'person aggrieved' under the statute to include 'legal representatives.'").
[80] *See Id.* at 23.
[81] *See* Miss. Code Ann. § 11-77-5.
[82] *Id.*
[83] *See* Bd. of Trs. v. Fox, 492 U.S. 469, 477 (1989).

This overbreadth analysis will focus on broader concepts of "material harmful to minors." While obscenity is not protected speech under the first amendment,[84] "material harmful to minors, the Mississippi law restricts, includes obscenity and material that is protected speech for adults, but still harmful to minors.[85] This restricted material beyond traditional obscenity's scope is grounds for an analysis on the constitutional overbreadth of the Mississippi law.

A successful overbreadth analysis renders the law "invalid in all its applications."[86] To challenge a statute on overbreadth grounds, the statute's overreach must be substantial in relation to the statute's legitimate scope, and the law must be the least-restrictive means of achieving the government's compelling interest.[87] An overbreadth argument is difficult to apply regarding commercial speech restrictions, as commercial speech is treated as more resilient to "chilling effects" that may occur from speech restrictions, when compared to noncommercial speech restrictions.[88]

The Mississippi statute restricting commercial entities distributing "material harmful to minors,"[89] will likely be considered commercial speech. Despite a large amount of the content being posted by individual users,[90] the commercial entities distributing this content are the ones subject to liability. And the distribution of content is for some commercial gain, transforming the restricted content into commercial speech.[91] But while the First and Fourteenth Amendments "protect commercial speech from unwarranted government intervention," the online restrictions of "material harmful to minors" are to an extent considered a compelling government interest, circumventing some of the protections attributed to this commercial speech.[92]

The identification requirement attached to the statutory age-verification could be an overbreadth of the statute's goals. It may create a chilling effect for adults who want to partake in the restricted material legally. The law requires adults to give personal identifying information to verify their age.[93] This may dissuade adults who would otherwise access this material, due to the fear of having to announce, even in an online setting, that they desire access to restricted content. By taking advantage of the adults' embarrassment, the Mississippi law creates a form of identity-based chilling, which may constitute overbreadth by the law in achieving its goals.[94] While the restrictions may be on commercial speech, the effects of the restrictions are felt by the website users. An adult, who would otherwise partake in the content, even if not

---

[84] Roth v. United States, 354 U.S. 476, 481 (1957).
[85] Ginsberg v. State of N. Y., 390 U.S. 629, 636 (1968).
[86] Bd. of Trs. v. Fox, 492 U.S. 469, 483 (1989).
[87] *See Id.* at 485.
[88] *Id.* at 481.
[89] Miss. Code Ann. § 11-77-5.
[90] *See How Many People are on Porn Sites Right Now? (Hint: It's a Lot.)*, FIGHT THE NEW DRUG, HTTPS://FIGHTTHENEWDRUG.ORG/BY-THE-NUMBERS-SEE-HOW-MANY-PEOPLE-ARE-WATCHING-PORN-TODAY/.
[91] Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n of New York, 447 U.S. 557, 562 (1980).
[92] *Id.* at 561; Romney, *supra* note 34 at 100 (citing Ashcroft v. ACLU, 542 U.S. 656, 675 (2004)) (holding that protecting minors from exposure to sexually explicit materials is a compelling government interest).
[93] Miss. Code Ann. § 11-77-3(h).
[94] *See* Bd. of Trs. v. Fox, 492 U.S. 469, 481 (1989).

obscene, may be unwilling to access the website in fear of losing their anonymity. Making the reach of the statute much wider than the statute proposes.

Additionally, the Mississippi statute is unclear on what databases must be used to verify the age of users.[95] This vagueness makes the statute's application difficult, as it is unknown how non-residents in the state will be affected. The databases used may not contain information on an international user, whose identification is from another country. Leading to situations where access is barred for an adult who should be able to access the content, if not for the age-verification systems required by the Mississippi statute.

This bar to users without acceptable forms of identification, may not be enough to challenge the law's constitutionality, if this obstacle is merely incidental to the state's compelling interest to restrict material harmful to minors. This requires a determination of if the Mississippi statute is the least-restrictive means of enforcement.[96]

The Mississippi statute creates obstacles for adults to access the restricted content as allowed by the statute. In the context of noncommercial speech, the obstacles of a chilling effect from providing identifying information in a traditionally anonymous space, combined with the required geolocation the commercial entities use to identify Mississippi users may be found to be more than incidental. But in this commercial speech context, these user obstacles to accessing restricted content is likely acceptable under an overbreadth analysis.[97] The saving-grace against an overbreadth challenge is the statute's threshold determination.[98] Addressing arguments that society is restricted to only material deemed suitable for minors.[99] Overall, the Mississippi statute will likely survive an overbreadth challenge to its constitutionality, as while its effects may be overbroad in certain situations, the threshold for unconstitutionality in an overbreadth challenge, especially for commercial speech, is very high. The statute may be constitutional in certain situations, depending on the content restricted, so an overbreadth argument would likely fail as the statute would not be "invalid in all its applications."[100]

C. Underinclusivity

The Mississippi law fails to address concerns with the imposed "reasonable" age-verification. The law creates inherent inequalities in treatment with its arbitrary thresholds and the vague language of the law itself. Underinclusive concerns, discussed in *Brown v. Entertainment Merchants Association,* may apply to the Mississippi law.[101]When dealing with first amendment rights, as the Mississippi law does here, they must be pursued by means that are not seriously underinclusive to be constitutional.[102]

---

[95] *See* Miss. Code Ann. § 11-77-3(h) (nothing is mentioned on the age-verification system's scope).
[96] Bd. of Trs. v. Fox, 492 U.S. 469, 478 (1989).
[97] *See Id.* at 477.
[98] Miss. Code Ann. § 11-77-3(i) (enacted) (Age-verification is not required for commercial entities whose platform is less than 33 and ⅓% "material harmful to minors").
[99] Butler v. State of Mich., 352 U.S. 380, 383-84 (1957).
[100] Bd. of Trs. v. Fox, 492 U.S. 469, 483 (1989).
[101] *See generally* Brown v. Ent. Merchants Ass'n, 564 U.S. 786 (2011).
[102] *Id.* at 805.

1. Issues with the Obscenity Threshold

The Mississippi law only restricts commercial entities whose websites are made of at least 33% of material harmful to minors, and does not require age-verification on websites that do not meet this threshold, even if those commercial entities distribute harmful material just at a lower ratio.[103] Allowing most social media platforms to operate unaffected by the Mississippi law, many of which distribute large amounts of harmful material online.[104] Allowing minors to access harmful material through these unrestricted avenues under the Mississippi law.[105]

This is underinclusive because it gives commercial entities the ability to escape liability, while distributing harmful material to minors online in Mississippi. This runs counter to the Mississippi statute's intended effect, which is to combat negative effects on the development of minors from early exposure to harmful material online.[106] By allowing some commercial entities to distribute this content without age-verification, and mandating the enforcement of age-verification systems on other commercial entities based on an arbitrary threshold, the statute is underinclusive due to an inequality in treatment. This is apparent when considering the threshold used is a ratio rather than an amount, meaning that some commercial entities affected by this law may distribute less overall "material harmful to minors" than a social media company, but the former entity will be subject to the law because their platform is smaller.

Additionally, the 33% threshold gives commercial entities a method to escape liability while distributing "material harmful to minors" without age-verification. The affected entities can inundate their websites with non-harmful material until the harmful material makes up less than a third of their website. Allowing websites to legally distribute harmful material without employing age-verification. The Mississippi statute does not prevent unaffected commercial entities from highlighting "material harmful to minors" present on their website, nor are there any bad faith considerations.[107] This inundation of content could be structure where the harmful material is easily found on the website, allowing any commercial entity to receive the same treatment as social media under this statute.[108]

This 33% threshold also creates unequal treatment of "material harmful to minors" online compared to in-person restrictions for this material. Rather than focusing on the restricted material itself, the law focuses on the entities distributing this content.[109] Different from how things are generally regulated in-person. For example, age-restricted content such as purchasing alcohol is regulated on the product, not the distributor, so nearly everyone verifies their age when purchasing the specific product,

---

[103] Miss. Code Ann. § 11-77-3.

[104] *See* Mike Wright, *Majority of teenagers 'now watching pornography on social media*', The Telegraph (May 5, 2021, 6:00 AM), https://www.telegraph.co.uk/news/2021/05/05/majority-teenagers-now-watching-pornography-social-media/#:~:text=The%20majority%20of%20teenagers%20are,messaging%20and%20social%20networking%20apps.

[105] *See Id.*

[106] *See* Miss. Code Ann. § 11-77-1.

[107] *Cf. Id.* at § 3 (The statute's language does not consider actions taken by commercial entities under the 33% threshold, and does not consider possible actions affected commercial entities could take to escape liability outside of direct compliance with the statute).

[108] *See* Alexander, *supra* note 7.

[109] Miss. Code Ann. § 11-77-5.

not when entering the store selling the product. If the framework of regulation applied to commercial entities by the Mississippi law were to occur in-person, then minors could purchase alcohol from Walmart because the majority of Walmart's sales are in non-harmful markets. If the Mississippi law's regulatory scheme would leave "material harmful to minors" accessible to minors without age-verification if applied in-person, then the Mississippi law is necessarily underinclusive for doing so online.[110]

This underinclusivity argument is supported by *Brown*, where the law at issue restricted violent video games, but not violence in Saturday morning cartoons, leaving children easy access to the subject-matter the law aimed to restrict.[111] The Mississippi law's 33% threshold does the same, as it only targets commercial entities distributing "material harmful to minors" whose platform is primarily made of this restricted content, while leaving other platforms distributing the harmful material the law aims to regulate free from restriction and accessible to minors. This underinclusive restriction raises doubts on the statute's true goal. Mississippi could be attempting to disfavor the business of certain commercial entities like pornography websites within the state, rather than legitimately pursuing the stated interest of restricting minors from accessing harmful material online.[112] Even if Mississippi is not maliciously targeting pornographic websites, the Mississippi statute is seriously underinclusive due to its exclusion of other commercial entities distributing harmful material online.

2. VPNs Ignored

Additionally, the Mississippi law disregards the potential use of VPNs by users to circumvent georestrictions put into place by commercial entities.[113] The law does not mention VPNs, and since liability is only possible for commercial entities, user actions seem irrelevant to the law's enforcement.[114] Meaning minors could still access restricted harmful material without age-restrictions blocking them, and the harmed individuals would have no recourse against the distributing commercial entity, if they are performing reasonable age-verification. As commercial entities are only forced to apply age-verification as outlined in the statute to users located in Mississippi. The best way to identify Mississippi users is through geolocation, but if the users located in Mississippi use a VPN to camouflage their physical location from geolocating technology, then users will not be subject to age-verification if they appear to be located in a different state or country.[115]

Commercial entities will only employ age-verification where it is required, and even if it is required in the location users pretend to be, the age-verification employed may be different than what is required by the Mississippi statute. Since the Mississippi law does not address this possibility, nor VPN use, then an individual harmed by accessing "material harmful to minors" would have no avenue to recover damages.[116] The Mississippi statute creates legal liability for commercial entities, while simultaneously restricting the efforts of individuals to recover damages through the statute. Because even if construed broadly, VPN companies do not meet the definition

---

[110] *See* Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 801-02 (2011).
[111] *See Id.*
[112] *See Id.* at 802.
[113] *See* Earle, *supra* note 48 at 11.
[114] *See* Miss. Code Ann. § 11-77-5.
[115] *See* Earle, *supra* note 48 at 11.
[116] *See* Miss. Code Ann. § 11-77-5.

of a commercial entity distributing "material harmful to minors" online.[117] Since VPN companies may enable individuals to access harmful material by circumventing the georestrictions for Mississippi, the VPN companies themselves do not meet the 33% threshold as they are a tool, and VPNs just provide their users a camouflaged IP address, they do not intentionally distribute harmful material to minors.[118] Leaving VPN companies and their users outside the Mississippi statute's scope.

VPNs are not a niche service that only a handful of individuals have access to. They are used by a large portion of the U.S. population for a variety of reasons, and the number of people using VPNs grows daily.[119] Many VPN services actively advertise their capability to hide a user's physical location from georestrictions, usually in the context of accessing alternate titles on streaming services.[120] Meaning VPN services are openly advertising a method for anyone in Mississippi to circumvent restrictions established by the statute. There are entire webpages dedicated to showing how to use a VPN to access pornographic material online.[121]

Laws restricting content online that do not account for VPN use are not narrowly tailored and are too underinclusive to be constitutional despite the government's compelling interest to protect minors from harmful content.[122] Because the Mississippi statute does not address VPNs, the law is underinclusively unconstitutional when compared to its asserted justification of restricting "material harmful to minors" online, which can include protected speech for adults.[123] This underinclusivity weakens the statute's justifications for restricting protected speech, even in a narrow manner, as the underinclusivity shows the statute can be further narrowed,[124] and currently the lack of VPN considerations likely makes the Mississippi statute unconstitutionally underinclusive.

3. Parental Veto

By placing the duty to act on a commercial entity's liability on the individual,[125] the Mississippi statute creates a "parental veto" in its enforcement. Since not all parents or individuals would pursue action against commercial entities because a minor accessed harmful material on their website.[126] If a parent allows their minor child to access this harmful material, then the law is only restricting the content to minors with parents prohibiting their child's access to harmful material online.[127] This "parental veto" shows the Mississippi statute is underinclusively unconstitutional because it is unequally applied and enforced.

Similar to how the violent video game restriction in the *Brown* case was deemed unconstitutionally underinclusive because a child can still access violent video games

---

[117] *Id.*
[118] *Id.* at § 11-77-3; Berry, *supra* note 5 at 488-89.
[119] Chauncey Crail, *VPN Statistics and Trends In 2023*, FORBES ADVISOR (Feb. 9, 2023, 12:51 AM), https://www.forbes.com/advisor/business/vpn-statistics/.
[120] Kyle Berry, *supra* note 5 at 489.
[121] *Se e.g.*, Lawrence Wachira, *supra* note 53.
[122] Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 802 (2011).
[123] *Id.*; Ginsberg v. State of N. Y., 390 U.S. 629, 636 (1968).
[124] *See* Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 802 (2011).
[125] Miss. Code Ann. § 11-77-5.
[126] *See* Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 802 (2011).
[127] *Id.*

so long as a single authority figure gives permission.[128] So, if a parent wants their children exploring "material harmful to minors" online, and gives them access to circumvention tools, like a VPN or their ID, then commercial entities distributing harmful content to those minors would not be liable if they enforce reasonable age-verification.

Since the Mississippi law's goal is to prevent minors from accessing "material harmful to minors" online,[129] then lack of state enforcement combined with the potential use of a "parental veto" would be underinclusive because it allows some children to access this material. It also permits noncomplying websites to escape liability, so long as minors who access them, have parental permission to consume potentially harmful material online. This "is not how one addresses a serious social problem,"[130] such as restricting access to "material harmful to minors" online. Currently, most commercial entities use an honor system as an age-verification method to prevent minors from accessing their websites. Unless the State shows that parents have a substantial need to prevent their children from accessing this harmful material, but are unable to do so, then the law's underinclusiveness may prevent the State's compelling interest from rising to the level of restricting constitutionally protected material online that is encompassed under "material harmful to minors."[131]

Overall, the various aspects of regulating the distribution of "material harmful to minors" online the Mississippi law fails to account for show the law's underinclusiveness. While the State may have a compelling interest in restricting this material, the statute's threshold determination, its failure to address VPNs, and the presence of a parental veto, harm Mississippi's justifications for restricting otherwise constitutionally protected material, likely making the statute itself unconstitutional on first amendment grounds for its underinclusive provisions.

D. Overinclusivity

The Mississippi law is also unconstitutionally overinclusive in what it restricts and the method it employs to do so. As stated in *Brown*, content restrictions must be narrowly tailored to serve the government's compelling interest.[132] The restriction of material harmful to minors is a compelling government interest, even though not everything restricted is obscene or harmful to adults.[133] But as society is not under the obligation of consisting of material only "fit for children,"[134] then restrictions on material that are only "variably obscene,"[135] must have strong justifications, or they are first amendment violations. The Mississippi law must likely survive a strict scrutiny analysis to be constitutional.[136] The following subsections show how the Mississippi law is not narrowly tailored to achieve its goal, and thus unconstitutionally overinclusive.

---

[128] *Id.*
[129] *See* Miss. Code Ann. §§ 11-77-1 & 11-77-5.
[130] Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 802 (2011).
[131] *See Id.* at 803.
[132] *Id.* at 799.
[133] Ginsberg v. State of N. Y., 390 U.S. 629, 640 (1968).
[134] Butler v. State of Mich., 352 U.S. 380, 383-84 (1957).
[135] Ginsberg v. State of N. Y., 390 U.S. 629, 635-366 (1968).
[136] *See* Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 805 (2011).

1. Disregard for Different Classes of Minors

The Mississippi law does not differentiate between different classes of minors below the age of eighteen.[137] By restricting all minors under the age of 18, the law is overinclusive. As the harms of the restricted content on minors, and the enforced age-verification may not apply nor be as effective for minors in the 16-18 age range.[138] This is especially prevalent in Mississippi where the age of consent is 16, rather than the national age of consent of 18.[139]

This lack of distinction supports the assertion the law is overinclusive and broad in relation to its inherent goal of preventing developmental harms in minors due to exposure to "material harmful to minors." Mississippi law allows 16-year-old minors to engage in sexual relations with older individuals, the effect of exposure to "material harmful to minors," would by comparison fall flat, and have a lower effect on these minors, compared to a young child exposed to harmful material. This difference in effect plus the similar restrictions faced by older and younger minors, hampers the first amendment rights of older minors. Older minors may not fall into the category of being affected by "material harmful to minors," especially if they can legally have intercourse with adults.[140] The Mississippi statute is being overinclusive in its restrictions, as older minors are being unnecessarily deprived of aspects of their first amendment rights as they would not be harmed by the restricted material similar to adult individuals.

The Mississippi law does not address whether emancipated minors would also be subject to restrictions, especially when emancipation occurs due to marriage.[141] The Mississippi law does not mention emancipation, so it can be assumed they will be subject to its age restrictions.[142] If so, then like older minors, the Mississippi statute is overinclusive as it restricts the first amendment rights of emancipated minors to "material harmful to minors." The law's restriction on this class of minors is unnecessary to achieve the State's compelling interest making the restriction unconstitutional.

2. Commercial Entity Response Effects

The Mississippi law addresses the liability of non-complying commercial entities, by allowing individuals to sue non-complying entities for damages minors experience due to this distribution of "material harmful to minors."[143] But the Mississippi statute does not account for malicious compliance complying entities, where instead of enforcing age-verification for users, the companies blocked all Mississippi users from accessing their websites.[144] The Mississippi statute is a

---

[137] *See* Miss. Code Ann. § 11-77-5 (the law applies to all individuals below under eighteen).
[138] *See Id.* at § 11-77-1 (Nothing in the statute's legislative findings indicate what age the negative effects found are most likely to occur when minors are exposed to harmful material.); Romney, *supra* note 34 at 72 & 103.
[139] *See* Miss. Code Ann. § 97-3-65.
[140] *See generally Id.* (If there are no supposed negative effects when a 16-year-old minor engages in sexual intercourse, then Mississippi's law is overinclusive for treating older minors as being just as in danger of harm from exposure to harmful material as younger minors).
[141] Miss. Code. Ann. § 93-11-65.
[142] Miss. Code Ann. § 11-77-5.
[143] *Id.*
[144] *See* Alexander, *supra* note 7.

restriction for minors, not adults.[145] This malicious compliance transforms an obstacle to access "material harmful to minors" for adults, into a ban.

The complete ban of Mississippi in response to this law may be considered overinclusive state action. As while no direct state action enforced these bans, they would never have occurred if not for this law, so even though commercial entities banned Mississippi as private actors, this is arguably a form of indirect state action.[146] Making this an overinclusive ban of "material harmful to minors" in effect, as the law allows companies to act this way, leaving adults with potentially no access to this material which is unconstitutional, as it could restrict online material available to just what is "fit for children."[147] Though content restrictions aimed at minors, are giver greater leeway, this total ban effect exceeds the reasonable allowance of broader restrictions.[148] This ban being a form of state action must be limited to protect the first amendment rights of adults in Mississippi. The statute needs account for malicious compliance to narrowly tailor the law as a restriction. Otherwise, the Mississippi statute should be reformulated as a ban for all users to encompass this malicious compliance.

3. Restrictions on Parental Autonomy

The parental veto discussed above may be used to argue the Mississippi law's overinclusivity[149] because some parents may believe allowing their children to access restricted material online is beneficial to the minor. By highlighting the harms of minors accessing "material harmful to minors,"[150] the State implies parents "ought" to stop their children and worry about their exposure to this material.[151] The Mississippi law is overinclusive for infringing on parental autonomy, rather than simply aiding in parental duties.[152]

Restriction on parental autonomy would occur in cases, where parents do not believe their child accessing what the law considers "material harmful to minors" to be harmful. Some Parent may even encourage their child to do so for sexual education. The Mississippi statute is taking the side of one type of parenting by restricting the content for minors completely, restricting the choice of parents on what values and methods they use to educate their children on subjects that are sexual in nature.[153] Even when protecting children's interests, constitutional limits on government action apply,[154] which should limit the Mississippi statute's application.

This alludes to the chilling effect of the law, which could render it unconstitutional if it stymies parental autonomy.[155] Some parents who want to sue a non-complying entity for damages, may not do so due to the embarrassment of the subject matter. The law creates this embarrassment, as it infers minors accessing this

---

145 *See* Miss. Code Ann. § 11-77-5.
146 *See* Davis, *supra* note 28 at 17.
147 Butler v. State of Mich., 352 U.S. 380, 383-84 (1957).
148 *See* Ginsberg v. State of N. Y., 390 U.S. 629, 640 (1968).
149 *See* Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 804 (2011).
150 Miss. Code Ann. § 11-77-1.
151 *See* Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 804 (2011).
152 *Id.*
153 *See Id.*
154 *Id.* at 804-05.
155 See *Id.* at 805.

material is wrong, and damages subsequently occur.[156] Leading to situations where parents may not want to identify themselves, as it may raise concerns on the parent's ability to parent, as their child accessed this harmful material despite their parental efforts. The Mississippi statute's stance on this issue, could overinclusively burden parental autonomy.

III. PUBLIC POLICY CONSIDERATIONS AND SOLUTIONS

While the above constitutional concerns are relevant, the Mississippi law and others like it could change to better limit the distribution of material harmful to minors online. The following are less restrictive and potentially more effective alternatives the current law. This will be explained through public policy considerations the law evokes. And suggested changes to the law to better enforce Mississippi's compelling interest in restricting minors' access to "material harmful to minors" online.

A. Broad Concerns

1. Ineffectiveness of the Law

A law's effectiveness has no strong bearing on any equal protection constitutional challenges it may face,[157] but addressing the Mississippi law's effectiveness is necessary when considering public policy that fueled its enactment. The legislative findings of the statute indicate the Mississippi law was enacted because of the harms the restricted material was found to have on minors' development.[158] The Mississippi statute framed itself as the answer to this public health concern.

If the Mississippi statute's goal is to prevent minors from accessing this material online, and not to restrict adults, then the ineffectiveness of the law should be addressed rather than be allowed to continue, so that it does not become perverse in its effect.[159] Otherwise, there are chances the effects of this law may turn out like mandated abstinence sexual education, where the law's effect became an obstacle, as instead of lowering rates pre-marital intercourse between minors, the abstinence education contributed to increased numbers in teen pregnancies and sexually transmitted diseases because of a lack of direct education.[160] Regarding the Mississippi statute, the number of minors exposed to harmful material online may increase because of age-verification, as it may push minors to find ways around restrictions rather than leaving them be. While complying websites are facing lower user circulation, there is no telling how much activity increased on non-complying websites.[161]

A counter-point is the Mississippi law may be ineffective towards minors actively seeking age-restricted material, but it is effective in preventing unknowing minors from finding restricted material accidentally.[162] This is potentially incorrect for

---

156 *See* Miss. Code Ann. § 11-77-1.
157 Railway Express Agency, Inc. v. New York, 336 U.S. 106, 110 (1949).
158 Miss. Code Ann. § 11-77-1.
159 Meghan Boone, ALSO FEATURING: Perverse & Irrational, 16 HARV. L. & POL'Y REV. 393, 409-10 (2022).
160 *Id.* at 430-434.
161 Brenna Goth, *Porn Site Age Checks Required by Growing Number of States*, BLOOMBERG L. (Jul. 26, 2023, 4:00 AM), https://news.bloomberglaw.com/in-house-counsel/porn-site-age-checks-required-by-growing-number-of-states.
162 Marsden*, supra* note 32 at 231.

two reasons. The first is while minors may stumble into restricted material accidentally, the presence of the law's age-verification may stoke the minors' curiosity. The age-verification's existence may compel minors to learn more about the subject because of the restrictions put into place. If this occurs, then minors may become part of the class of minors actively seeking age-restricted content online. If this occurs then the law's ineffectiveness, due to the ease in circumvention and many non-complying commercial entities, will come into play, allowing minors to be harmed by the material the statute aims to restrict. Since the Mississippi statute is so new, its effectiveness compared to the problematic honor system currently in use is unknown,[163] but it may not be an adequate replacement, especially if there is no incentive to increase compliance with the law.

The second issue with this accidental prevention argument, is if the goal of the legislature was only to prevent younger minors from accidentally discovering "material harmful to minors" then the statute would have been written reflecting that goal. Instead, the statute is written to prevent the access of all minors, and places a duty on certain commercial entities.[164] If the law wanted to only prevent accidental access to this material, then the statute should have included a liability exception for entities whose age-verification was actively bypassed by a user.[165] The Mississippi law creates potential liability for commercial entities, and disregards any potential user liability.[166] This lack of addressing user circumvention, shows the Mississippi statute's goal is to prevent all minors from, accidentally or purposefully, accessing "material harmful to minors" online. Meaning when minors circumvent the age-verification systems of commercial entities, under the current law, there is no recourse for damages because the commercial entity complied with the law. Creating scenarios where the law's ineffectiveness allows minors to legally experience harms the law was created to prevent.[167] If this is the stance the State wants to take on the issue of distributing "material harmful to minors" online, then it should try preventing these harms from occurring more often than not.

The lack of policing method to monitor non-complying commercial entities diminishes the law's effectiveness. While it need not be perfect, if the most common results from online searches of subject under the "material harmful to minors" umbrella do not consist of complying commercial entities, then the law is so ineffective, that it is inconsequential in its effect and enforcement.[168] This non-compliance counters arguments for the Mississippi law's enforcement that since many common activities require valid identification, then verifying age to access this harmful material should be treated similarly.[169] Because if an individual is never asked to reasonably verify

---

[163] See *Id.* at 227.
[164] Miss. Code Ann. § 11-77-5.
[165] *See Id.*
[166] *See Id.*
[167] *See Id.* at § 11-77-1.
[168] This standard is difficult to determine as search results for the same search terms vary based on factors such as time, user location, and the user's past searches, but if non-complying websites appear more often due to online porn restrictions, then these laws' effectiveness should be further analyzed. *See generally Why your Google Search results might differ from other people*, GOOGLE ((last visited Dec. 12, 2023), https://support.google.com/websearch/answer/12412910?hl=en.
[169] Marsden, *supra* note 32 at 239 ("The ultimate response is that valid forms of ID are required for many day-to-day activities, including driving, purchasing alcohol, voting, and even going to the movies. Therefore, it is also reasonable to require those who access pornography to show that they are legally permitted to consume it.").

their age, due to non-compliance, then the law is ineffective since the State cannot make entities comply. Especially for the distribution of harmful material online, as much of it is available for free, and not subject to age-verification at the "point of delivery," or pay walls like when purchasing alcohol, meaning minors can access the harmful material immediately if there is no age-verification.[170]

It is unclear if the State can enforce fines on non-complying entities, but some sources report that it will occur.[171] The threat of fines or lawsuits against non-complying commercial entities are less effective for those based outside of the United States, as they are not subject to much "legal jeopardy" compared to domestically based commercial entities.[172] Even if fines becomes the predominant enforcement mechanism of the Mississippi statute, "if the law's enforcement is so rare that the lawmaker's coercive intent is not translated to non-complying commercial entities to alter their behavior, then the law itself is as coercive as a parent's rules to their children, which depend on the parent to follow through with their enforcement."[173] The Mississippi law only seems to be effective once damages are sought after the fact, and if law's goal is preventative, then it needs to be more coercive to alter the behavior of non-complying commercial entities to prevent minors from accessing "material harmful to minors" online.[174]

2. Implications of a National Trend

This Mississippi law is indicative of a national trend towards stricter age-verification online. The Mississippi statute is one in a trend of other states, like Louisiana and Virginia, in establishing this type of online restriction for minors.[175] The issues in the Mississippi law, both in its constitutionality and effectiveness, must be evaluated with greater scrutiny, as this trend shows this law is not an outlier, but part of a new movement. Showing a willingness from states to create legislation using geoblocking to restrict online content. Before this trend of states action, georestricting was mainly used to enforce licensing agreements between companies.[176]

This growing trend of states enforcing georestriction-based age-verification systems on commercial entities, like the Mississippi statute,[177] increases the importance of analyzing the method of restricting content online these laws employ. This trend is especially apparent when comparing the language of the laws themselves, many of which use differing terms, and may not encompass all of the same material.[178] If the number of states employing this method of age-verification continues to expand,

---

[170] *See* Jessica Muirhead, *Preventing underage alcohol purchasing online using payment card details*, INST. OF ALCOHOL STUD. 1 , 7-10 (Dec. 2021), https://www.ias.org.uk/wp-content/uploads/2021/12/IAS-Preventing-underage-alcohol-purchasing-online-using-payment-card-details.pdf.

[171] *See* Miss. Code Ann. § 11-77-5; *see e.g.*, Alexander, *supra* note 7. (reporting "Senate Bill 2346 became official … websites considered pornographic or "obscene" must now have strict age verification processes or face a fine from the attorney general.").

[172] *See* Mcintyre, *supra* note 60.

[173] Joseph D'Agostino, *Law's Necessary Violence*, 22 TEX. REV. LAW & POL. 121, 182-83 (2017).

[174] *See Id.*

[175] *See* Goth, *supra* note 161.

[176] *See* Earle, *supra* note 48 at 11.

[177] *See* Miss. Code Ann. §§ 11-77-1 – 11-77-7.

[178] *Compare id.* at § 11-77-5, *with* VA Code Ann. § 18.2-391 (Virginia's version of Mississippi's law), *and* A.C.A. § 4-88-1305 (Arkansas' version of Mississippi's law).

then the Mississippi law's issues outlined above, may be exacerbated, and states that do not mandate age-verification may be affected by laws enforced in other states. And rather than allow these issues to continue and worsen, it would be effective long-term, if states like Mississippi, who are employing restrictions on "material harmful to minors" online, address the statute's issues while the law's effect is relatively minimal. Even if only one state addresses these issues, it would serve as a blueprint for others who have enacted or are seeking to enact a similar law.

Outside of this "material harmful to minors" context, if these laws survive constitutional scrutiny, then they could be the foundation for how content is regulated online. The age-verification systems required by the Mississippi statute could be applied to things like online gambling and the sale of alcohol. Affecting much broader markets than just obscene or harmful material for minors. It is unknown what will be restricted online in the future as technology progresses, but addressing inefficiencies or constitutional concerns in the Mississippi law now, would ease the creation of future online restrictions.

3. International Solutions

Looking at how other nations regulate harmful material online, provides potential alternatives to the Mississippi law that could better survive constitutional scrutiny, or evoke compliance from commercial entities.

A feasible option is the "porn pass" distributed in the United Kingdom.[179] This is a physical form of age-verification, where users go to a store, show the clerk their identification and obtain a physical card, allowing them to access restricted online content.[180] This would treat accessing restricted content online similarly to purchasing alcohol in-person. Chilling effects of providing identifiable information would be reduced, as only a single clerk has to verify the user's age, in-person, with little chance of stealing their information compared to online age-verification. By removing online age-verification, many privacy concerns will be put to rest.[181] When accessing a restricted website, the user inputs the relevant information from their "porn pass" rather than providing websites identifying information directly. Classifying access to material harmful to minors in the same "day-to-day activities" group requiring valid identification to partake in.[182]

Alternatively, Germany regulates a self-regulating body for online content.[183] This self-regulating body is made of member organizations creating rules for members to follow when restricting content, based upon governmental guidelines.[184] Companies join this self-regulating organization to optimize "youth protection online," and to give commercial entities a say in the regulations' form, creating a method of "voluntary self-regulation."[185] The self-regulating body enacts and enforces online restrictions, while the government regulates the self-regulating body, instead of individual commercial

---

[179] Romney, *supra* note 34 at 69.
[180] *Id.*
[181] *See* Alexander, *supra* note 7.
[182] *See* Marsden, *supra* note 32 at 239.
[183] Romney, *supra* note 34 at 76.
[184] *Id.* at 77.
[185] *See* FREIWILLIGE SELBSTKONTROLLE MULTIMEDIA-DIENSTEANBIETER, https://www.fsm.de/en/fsm/ (last visited Nov. 11, 2023).

entities.[186] The government issues sanctions and takes legal action if commercial entities are violating laws on the dissemination of harmful online content, but if the self-regulating body is "acting within the scope of its discretionary powers," and its members comply with its regulations, then the government is not to discipline individual entities.[187] If applied to Mississippi's law, it could lead to greater compliance from affected commercial entities, as they will have an inputs on the regulation's form. This requires greater resource investments, than just enacting the law, but could be useful in increasing the efficacy of age-verification, and commercial entity compliance. Creating clearer avenues to recover damages, as non-complying entities are distributors of "material harmful to minors," and are monitored by the self-regulating body.[188]

The above two alternatives to the Mississippi law, show the goal of restricting online content is popular, but the method employed by Mississippi is not the only way to achieve it.

B. The Mississippi Law's Problematic Provisions

1. Issues with the Law's "Serious Value Exception"

The Mississippi law serves as a check and a guide for commercial entities of the law's effect on them by defining "material harmful to minors." The statute's serious value exception prevents restrictions of "material harmful to minors" if the work taken as a whole has "serious literary, artistic, political, or scientific value for minors."[189] The serious value exception, which mirrors the language in *Miller* but adds a focus for minors,[190] is inadequate as a guide for commercial entities due to difficulty in determining what harmful material has serious value. This is especially true for "material harmful to minors," which is broader than obscenity as defined in *Miller*.[191] To serve as a preventative guide for commercial entities, the law must clarify whether certain categories of harmful material have serious value. This distinction does not have to encompass everything, but the law should define more common and emerging forms of online harmful material not contemplated when *Miller* was decided in 1973.[192]

The serious value exception's scope for "material harmful to minors" must be clarified. Currently harmful "material taken as a whole which lacks serious literary, artistic, political, or scientific value for minors" is restricted by the Mississippi law, and counts towards the 33% threshold where commercial entities are subject to the law.[193] This definition excludes harmful and obscene material if they have serious value from this threshold determination. Courts subject harmful material to a balancing test to evaluate serious value, but this ambiguity of what has serious value leaves companies

---

[186] *See* Romney, *supra* note 34 at 77-78.
[187] *Id.* at 77.
[188] *See* Miss. Code Ann. §§ 11-77-3 – 11-77-5.
[189] Miss. Code Ann. § 11-77-3.
[190] *See generally* Miller v. California, 413 U.S. 15 (1973); *see also* Miss. Code Ann. § 11-77-5.
[191] *Id.*; *See generally* Miller v. California, 413 U.S. 15 (1973) (the test for determining obscenity, does not apply to everything the Mississippi law could restrict).
[192] *See generally id.*
[193] Miss. Code Ann. § 11-77-3.

unaware if the material they distribute has serious value and excluded from the Mississippi law.

The law should classify categories of obscene material, that may have serious value, and clearly restrict them to avoid their availability online to minors. Or specifically exclude them from restrictions depending on the legislature's stance. This would address large categories of otherwise harmful material, excluding or including commercial entities specializing in their distribution from the Mississippi statute. Examples of harmful material with arguable value that should be addressed include: deepfake pornography (scientific value), obscene drawings (artistic value), and artificially generated images (artistic and scientific value). These all include harmful material to minors that are found online with no age-verification. The Mississippi law should address whether the above categories generally qualify for the serious value exception.[194] This determination would expand the effectiveness and the scope of the Mississippi statute for users and commercial entities who would be affected by the law depending on if this material lacks has serious value.

Deepfakes are altered videos, where deepfake technology puts the faces and expressions of others, like celebrities, onto the bodies of other people in videos.[195] This can be obscene, as many have put the faces of celebrities onto performers in explicit videos online. [196] Do the learning opportunities of this technology give it serious value here? Some say yes, as the "benefits of deepfakes' underlying technology" allow improvement of the technology underlying automated systems.[197] "Proponents for deepfake protection argue that any restrictions on deepfakes," even pornographic ones, "would have a chilling effect on deepfake technology," deepfake technology developers would fear the "possibility of facing a lawsuit."[198] This chilling effect may impact progress and development of new technology as outlined in the Constitution's Patents and Copyright section.[199] Others argue deepfake obscenity does not have serious value because they are "falsely depicting someone in pornography. Even people who create pornographic deepfakes acknowledge that what they do is derogatory."[200]

This debate's existence shows the serious value of deepfakes is unclear. Providing commercial entities distributing deepfakes plausible deniability for not instituting age-verification, as they can argue deepfakes, obscene or not, have serious technological value.[201] This vagueness in serious value determinations could be applied and argued to any of the above harmful material categories. Until case law, or the law itself, addresses this vagueness, commercial entities will distribute this harmful material to minors online without age-verification. Addressing the serious value exception's scope for these categories would allow commercial entities to evaluate the statute's applicability to them increasing their compliance.

---

194 See *id.*
195 Waldstreicher, *supra* note 23 at 731-33.
196 *Id.* at 733-34.
197 *Id.* at 756.
198 *Id.*
199 *See* U.S. CONST. art. I, § 8, cl. 8.
200 Waldstreicher, *supra* note 23 at 755-56.
201 See *Id.* at 756.

2. Alter the 33% Threshold

An alteration to the Mississippi statute could be changing the threshold requirement determining the commercial entities affected. The Mississippi statute affects commercial entities whose websites are at least 33 % "material harmful to minors."[202] This threshold leaves much "material harmful to minors" available online without age-verification, and gives commercial entities a method of escaping liability by lowering the ratio of harmful material on their website to under 33%.[203]

The Mississippi law could be altered into a post-by-post restriction that directly targets "material harmful to minors." Where all commercial entities would statutorily age-restrict any content on their websites that qualifies as "material harmful to minors." Functioning similar to how YouTube age-restricts its content.[204] YouTube identifies elements of a post that are unsuitable for viewers under 18, and age-restricts specific posts with these elements.[205] Once age-restricted, the post becomes unviewable for users unless they log into their YouTube account which has verified their age as 18 or older.[206]

Mississippi could adopt YouTube's age-restriction model, and require all commercial entities age-restrict any harmful material content on their websites.[207] The mandated age-verification, which requires valid identification,[208] could be instituted when users create their accounts, rather than when users access the website. Commercial entities could then mark specific accounts as age-verified, and provide users of those accounts access to restricted content, when logged into their age-verified account, allowing any non-verified user access to non-harmful material on website.

This restriction would apply to entities distributing "material harmful to minors" online, enforcing age-verification for all restricted content distributed online, not just websites with larger concentrations of this content.[209] Allow restrictions to occur on social media, where many minors are accessing harmful material, and other websites unrestricted under the Mississippi law.[210] Age-verification on account creation would prevent minors from creating age-verified accounts, while still providing access to the non-harmful materials on websites that are currently restricted. Commercial entities would have to flag and age-restrict content, similar YouTube, on their platform,[211] tailored to "material harmful to minors."[212]

---

202 Miss. Code Ann. §§ 11-77-3 – 11-77-5.
203 *See* Section III.C.1.
204 *See Age-Restricted Content*, YOUTUBE HELP, (last visited Nov. 11, 2023), https://support.google.com/youtube/answer/2802167?hl=en (explaining YouTube age-restricts specific posts based on community guidelines and terms of service.).
205 *Id.*
206 *Id.*
207 *See* Miss. Code Ann. § 11-77-3.
208 *See id.*
209 *See id.* at §§ 11-77-3 – 11-77-5 (age-verification is currently only enforced on entities whose websites are made of greater than 33% "material harmful to minors" regardless of the website's size).
210 *See* Wright, *supra* note 104.
211 *See Age-Restricted Content*, *supra* note 204.
212 *See* Miss. Code Ann. § 11-77-3.

C. Technological Concerns

The Mississippi law and others like it highlight technological considerations that must be deliberated when enforcing online regulations. The law's constitutionality and effectiveness can hinge on current technological limitations of technology used to enforce legal mandates online, especially where a human presence is not present to enforce the law.[213]

1. Current Technological Limitations

Many issues with the Mississippi law because of limits in technology used for age-verification and geoblocking. To employ age-verification for Mississippi users, commercial entities must employ geolocation technology for companies to identify a user's location, to detect and restrict Mississippi users, while allowing users from non-restricting states to access their website as usual.[214] The biggest obstacle to effective geoblocking is the IPv4 system.[215] Under this IP system, the geolocation accuracy and subsequent restrictions are not 100% accurate, especially when identifying user's specific state, as user geolocation is about 50-80% accurate when determining the user's state within the country.[216] Meaning commercial entities may accidentally restrict non-Mississippi users in surrounding states, due to geolocation inaccuracy, when enforcing the Mississippi law, or vice versa, where Mississippi users are flagged as from different state, and allowed unrestricted access to websites age-restricting Mississippi users. Which is worse when compared to geolocation accuracy of 95-99% when distinguishing different countries.[217]

When devices transition to the IPv6 system, then of the law's restrictions will become easier to enforce due to device specific IP addresses making it easier to detect when VPNs were used.[218] As under IPv6, each device would have its own static IP address, due to larger numbers of available IP addresses, rather than randomly assigned addresses, used by anyone, anywhere.[219] Implementation of IPv6 is not complete, and less than 50% of devices have implemented this new system.[220]

Under IPv4, age-verification required by the Mississippi law is circumventable to the point where its enforcement may be unconstitutionally broad with current technology. In the past, the Supreme Court has focused on relating the constitutionality of online content restrictions to the limits of the technology performing the restrictions as narrowly defined by the law to not be unconstitutionally burdensome.[221] The lack

---

[213] *See* HOLMES, *supra* note 2 at 14; *see e.g.*, Reno v. ACLU, 521 U.S. 844 (1997).
[214] *See* Earle, *supra* note 48 at 7.
[215] Trimble I, *supra* note 35 at 595 (All conceivable IP addresses under the "IPv4 protocol have been assigned, internet service providers assign and reassign IP addresses from a common pool of them to internet users as they log in and off from online services," making it hard to track where an its user is actually located).
[216] Emma Jagger, *Why IP Geolocation Can Go Wrong: Causes and Fixes*, ABSTRACT API (Aug. 4, 2023), https://www.abstractapi.com/guides/why-is-my-ip-geolocation-wrong.
[217] *Id.*
[218] See Trimble I, *supra* note 35 at 595-97.
[219] *Id.*
[220] Josh Fruhlinger, *What is IPv6, and why is adoption taking so long?*, NETWORK WORLD (Mar. 21, 2022, 3:00 AM), https://www.networkworld.com/article/3254575/what-is-ipv6-and-why-aren-t-we-there-yet.html.
[221] HOLMES, *supra* note 2 at 14.

of available technology preventing the access of only minors from restricted content was why the CDA was repealed, while the existence of blocking technology already in use led the Court to concluding other less restrictive means of achieving the government's compelling interest exist.[222] Meaning if technology is not sufficiently advanced then the Mississippi law risks overboard enforcement.[223]So the Mississippi statute may not currently be constitutional, but when IPv6 is implemented the Mississippi statute could be applied effectively, narrowly, and within the constitutional scope for online content restrictions.

2. Address VPN Use

The Mississippi law needs to address VPN use. If law makers would prefer keeping user liability outside the law's scope, then making VPN companies liable for aiding user georestriction circumvention would address significant underinclusive arguments.[224] VPNs are used by consumers to trick georestrictions enforced by companies, to access copyrighted or restricted material in their area.[225] This ability to change online locations to circumvent georestrictions is advertised by VPN companies as a major feature of their service.[226]

Geotraveling can circumvent age-verification on websites complying with these "porn" laws, like the Mississippi statute, allowing minors to access harmful material online, and currently VPN companies are not liable these laws, as they do not meet their threshold requirements.[227] The presence of VPNs limits these laws' restriction to only minors who cannot access VPNs to circumvent age-verification.[228] VPNs allow technologically adept, usually older, minors to circumvent instituted age-verification,[229] which is counter to the Mississippi law's goal.[230] If the Mississippi law only wanted to restrict accidental exposure of younger minors to harmful material, then it would not have highlighted the negative effects pornography has on adolescents, by limiting the findings to only its effects on pre-pubescent minors.[231]

The Mississippi law could enforce age-verification to access VPN services, similar to how it restricts websites distributing "material harmful to minors."[232] It would be difficult to outright ban VPNs as they are used for privacy and security by individual users and companies.[233] Meaning age-verification for VPNs should be

---

[222] *Id.* (referencing the holdings in Reno v. ACLU, 521 U.S. 844, (1998) and Ashcroft v. American Civil Liberties Union (Ashcroft II) 1 542 U.S. 656 (2004)).
[223] *See id.*
[224] *See* Section III.C.2; *see* Michelle Edelman, *The Thrill of Anticipation: Why the Circumvention of Geoblocks Should be Illegal*, 15 VA. Sports & Ent. L.J. 110, 129 (2015).
[225] *Id.* at 116.
[226] *Id.* at 120.
[227] *See* Wachira, *supra* note 53 ("Laws and regulations regarding … porn vary significantly from region to region, while some local networks restrict access to porn sites…. A VPN is the easiest and most reliable way to get around these restrictions. This simple app changes your virtual location …. That way, you'll appear as if you're in another country that doesn't restrict access to porn sites."); *see also* Miss. Code Ann. § 11-77-3.
[228] Romney, *supra* note 34 at 72.
[229] *Id.*
[230] *See* Miss. Code Ann. §§ 11-77-3 – 11-77-5 (restricting minors under the age of 18 from harmful websites through age-verification).
[231] *See Id.* at § 11-77-1.
[232] *See Id.* at § 11-77-5.
[233] *Best VPNs of September 2023*, *supra* note 53.

limited to the cybertraveling features, rather than any security or privacy features the service provides. Allowing VPNs' cybertraveling feature to be used for "legitimate purposes" rather than for actively circumventing state laws.[234]

This would hold VPN services accountable for helping minors circumvent age-verification, even if this circumvention is not intended by the VPN companies. Entities currently restricted by the Mississippi law will not know if users circumventing their age-verification systems are minor, so this layer of age-verification on VPNs should reduce the number of minors able to circumvent these age-verification systems. Categorizing the use of VPNs to alter geolocation to access "material harmful to minors,"[235] with activities requiring valid identification to partake in.[236]

Others argue VPNs are not foolproof loopholes to age-verification because there is technology that detects VPN use and blocks commonly used IP addresses for geotraveling, preventing their circumvention of age-verification.[237] This technology is used by services like Netflix to enforce georestrictions, but are circumvented by more expensive VPNs.[238] Disregarding feasibility or expense issues related to enacting VPN detection systems, there is currently no incentive for commercial entities to use them, since the Mississippi law does not address VPNs. Commercial entities that enacted age-verification solely to comply with Mississippi's law will not use VPN detection systems if they are not mandated by the law. This desire not to act beyond the minimum mandates can be seen from the malicious compliance commercial entities have performed in response to Mississippi's law.[239]

If restricting access to VPNs would be difficult, then having VPN services give their users notice may be sufficient. This could be a disclaimer by VPN services on their interface, informing users that using VPNs to circumvent age-verification may have legal ramifications. This may not prevent VPN users from circumventing georestrictions, but the disclaimer should give parents, who may be unaware VPNs can bypass age-verification, notice.[240] Allowing parents to better monitor their children's online activities.

3. VPNs in the American Legal System

The Mississippi law and its failure to address VPNs highlights a tendency in the American legal system to ignore VPNs.[241] Reluctance to address VPNs occurs, not only for restrictions on "material harmful to minors" online, but VPNs are also ignored

---

[234] *See* Trimble I, *supra* note 35 at 648-49.
[235] Miss. Code Ann. § 11-77-5.
[236] *See* Marsden, *supra* note 32 at 239 (positing that many "day-to-day activities" require valid ID to partake in, and pornography should be treated the same way).
[237] *Id.* at 238-239 (*citing Frequently Asked Questions for Clients,* AGE VERIFICATION PROVIDERS ASS'N https://avpassociation.com/av-clients/faqs-for-clients/ (last visited Mar. 29, 2023)).
[238] *Id.*
[239] *See* Alexander, *supra* note 7 ("Pornhub — one of the largest and most well-known adult content websites in the U.S. — has banned Mississippi users from accessing its content in response" to the Mississippi law, rather than enforce the statute's age-verification on users.).
[240] *See* Marsden, *supra* note 32 at 212.
[241] *See e.g.* Miss. Code Ann. § 11-77-5.

in fields like online gambling[242] and the DMCA.[243] This reluctance to address VPN use, even though they create complications in multiple legal fields, is odd. This could indicate these online restrictions are forms of political theater, and lawmakers do not care to effectively regulate VPNs, so long as they receive political credit for enacting hot topic laws like online porn restrictions for minors.[244]

While VPNs have legitimate uses,[245] they can be used for illegitimate uses like violating user agreements and performing illegal acts online gambling, if users appear as being from a different location geographically.[246] This lack of laws addressing VPNs creates legal ambiguity, where VPN services are advertised as legitimate, despite facilitating the illegitimate acts of their users.[247] Users are unaware if their use of VPNs to circumvent georestrictions is an issue because the laws they are breaking never contemplated the legal ramifications of VPN use.[248]

MGM Studios Inc. v. Grokster, Ltd., provides a potential framework for holding VPN companies accountable, at least regarding copyright infringement.[249] Holding that distributors of a device, or service in this case, promoting copyright infringing uses of their product may be secondarily liable for the direct infringement of third parties using that product, as the potential infringing use of a product alone is insufficient.[250] VPN companies actively advertise, usually through sponsorships, how their geolocation services can circumvent georestrictions in a user's area, encouraging the use of their VPN to new customers to aid in circumventing georestrictions used by

---

[242] *See* MISSISSIPPI GAMING COMMISSION, *Frequently Asked Questions*, MS. GAMING COMM'N, https://www.msgamingcommission.com/faqs#:~:text=internet%20gambling%20legal%3F-,No.,from%20Mississippi%20with%20these%20businesses (last visited Nov. 17, 2023) ("Internet gambling is illegal under state law. Online sites may advertise they are 'legal' and 'licensed' forms of gaming. They may be legal or licensed where the bets are received, but it is illegal to place bets from Mississippi with these businesses."); *see* Miss. Code. Ann. § 97-33-1 (no mention of VPNs or circumventing restrictions on online gambling in Mississippi).

[243] *See* 17 U.S.C. § 1201 (Digital Millenium Copyright Act ("DMCA") section covering circumvention of copyright protection systems.); *see* Berry, *supra* note 5 at 517 (circuit split in the United States on if the act of circumvention "is sufficient for liability or whether the act of circumvention must be connected to an act of infringement.").

[244] This lack of VPN addressal shows a tendency in politics to portray complex matters into simple ones, to show constituents that change is being made, without regard to the details of the change itself. *See* Kenneth L. Karst, *Faiths, Flags, And Family Values: The Constitution of The Theater State*, 41 UCLA L. REV. 1, 5 (1993). This political theater is shown in how media has covered the enactment of these online porn restrictions, where the broad effects of the law, the bipartisan support, and the politicians taking credit for their enactment are highlighted, and the possibility of circumvention through VPNs is barely covered. *See e.g.*, Jackson, *supra* note 56. If VPNs are mentioned in media coverage or by politicians, they are treated as inconsequential workarounds. *See e.g.*, Novicoff, *supra* note 3.

[245] *See Best VPNs of September 2023*, *supra* note 53 (listing legitimate functions of VPNs for personal and business use).

[246] *See* Berry, *supra* note 5 at 488-89.

[247] *See* Crail, *supra* note 119 (outlining predominant uses for VPNs, while advertising three VPN services that perform all those services as legitimate.).

[248] *See e.g.*, Miss. Code Ann. § 11-77-5; *see also* Miss. Code. Ann. § 97-33-1.

[249] *See generally* MGM Studios Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

[250] *Id.* at 941.

streaming services to access content available elsewhere.[251] Applying *Grokster* to VPNs is possible to induce secondary infringement because of their advertisements.

Outside of a copyright context, while *Grokster* may not apply, its reasoning could be applied to hold VPNs accountable in other legal contexts.[252] VPNs do not directly advertise that users should circumvent age-restrictions, but the infringing use these companies do advertise is so closely related to age-verification circumvention that VPNs should be held accountable. At least to inform lawmakers, so they can address VPNs in relation to laws like those restricting online gambling, and mandating age-verification online.[253]

Restricting VPN use may be difficult due to its virtual nature, and that many VPN companies are based outside of the United States,[254] but that has not stopped lawmakers from restricting content online.[255] Lawmakers should incorporate, at minimum, some form of notice regarding VPNs circumventing online restrictions. Allowing VPN users to be aware that using VPNs to circumvent georestrictions may lead to violating user agreements,[256] or potentially subject them to litigation if used to circumvent legally mandated online restrictions.[257] This notice could be the first step needed to begin addressing VPN use in the American legal system, as the need for stronger legislation grows with the popularity of VPNs.

CONCLUSION

Overall, this wave of restrictions on the distribution of "material harmful to minors" online through enforcement of stricter age-verification systems has serious constitutional and public policy implications. While the laws' constitutionality is debatable under the first amendment, the laws highlight how their enforcement mechanisms may be problematic in achieving their overarching goal, which is restricting minors from accessing harmful material online.[258] These laws show that

---

[251] *See e.g.*, NORDVPN, *What is a VPN and how it works | NordVPN*, YOUTUBE (Sep. 14, 2020), https://www.youtube.com/watch?v=yCWNRzoQGis (video posted on YouTube by a large VPN company, where one VPN feature is accessing blocked content.); Globku, *Ranking Every Naruto Storm Connections Ultimate*, YOUTUBE (Nov. 18, 2023), https://www.youtube.com/watch?v=aFzrAQNjsQs (YouTube video sponsored by a VPN company, with an in-video advertisement, highlighting the use of the VPN to access Netflix libraries in other countries from minutes 1:53-3:16).

[252] *See generally* MGM Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 939-41 (2005) (the companies at issue were actively advertising how their users could infringe copyrights, and the entities selling the service profited off infringing uses of their service.).

[253] *See e.g.*, Miss. Code Ann. § 11-77-5; *see also* Miss. Code. Ann. § 97-33-1.

[254] Dovydas Vėsa, *Who owns your VPN? 105 VPNs run by just 24 companies*, VPN PRO (Aug. 10, 2023), https://vpnpro.com/blog/hidden-vpn-owners-unveiled-97-vpns-23-companies/ (many VPNs are based outside of the United States in countries such as China, Pakistan, and Panama).

[255] *See* Miss. Code Ann. § 11-77-5 (restricting commercial entities distributing "material harmful to minors" online, without regard to where the commercial entities are based).

[256] *See Netflix Terms of Service*, NETFLIX (last updated Jan. 5, 2023), https://help.netflix.com/legal/termsofuse (Netflix's terms of service prohibit users from "**circumventing**, removing, altering, deactivating, degrading, blocking, obscuring or thwarting any of the content protections … of the Netflix service, including … copyright notices, and trademarks." (emphasis added)). This circumvention includes using VPNs, as if detected, Netflix will restrict users' access to only content that Netflix holds a worldwide license to stream. *Watching TV shows and movies through a VPN*, NETFLIX, https://help.netflix.com/en/node/114701.

[257] *See* Miss. Code. Ann. § 97-33-1 (violating Mississippi's anti-gambling law can lead to a fine of up to $500 and potentially up to 90 days of imprisonment.).

[258] *See* Miss. Code Ann. §§ 11-77-1 – 11-77-7.

despite VPNs' prevalence in the market, and their potential to circumvent online restrictions, that lawmakers tend to ignore their effect in undermining these laws' effectiveness. If state enforced online restrictions, like Mississippi's law, become the norm in the United States, then this lack of contemplation on VPNs' place in society will become an issue when enforcing these restrictions.

The online nature and multi-state push to restrict content online, may indicate this as an issue Congress should address, rather than leaving it to the states. Primarily as regulation would be easier on a national level for the government and the commercial entities. If the federal government passed a modern version of Mississippi's law nationally, then georestricting would be more accurate,[259] and commercial entities would only have to follow one set of rules rather than multiple varied sets of rules existing between each state's version the law.[260]

A national version of these laws could serve as the foundation for future laws restricting content online outside and within the context of material harmful to minors, and if the national law addressed VPN use, then lawmakers may begin to address VPNs in a direct manner that has yet to occur. As the current state of ignoring VPNs will not suffice going forward. The necessity of legal blueprints addressing VPNs is forming in our increasingly online world.

---

[259] *See* Jagger, *supra* note 216 (IP geolocation between countries is 95-99% accurate, whereas IP geolocation between states is 55-80% accurate.).

[260] *Compare* Miss. Code Ann. §§ 11-77-1 – 11-77-7, *and* A.C.A. § 4-88-1305*, with* La. R.S. § 51:2121.

THE EXPLANATIONS ONE NEEDS FOR THE EXPLANATIONS ONE GIVES—THE NECESSITY OF EXPLAINABLE AI (XAI) FOR CAUSAL EXPLANATIONS OF AI-RELATED HARM: DECONSTRUCTING THE 'REFUGE OF IGNORANCE' IN THE EU'S AI LIABILITY REGULATION

Ljupcho Grozdanovski*

Abstract: This paper examines how explanations related to the adverse outcomes of Artificial Intelligence (AI) contribute to the development of causal evidentiary explanations in disputes surrounding AI liability. The study employs a dual approach: first, it analyzes the emerging global caselaw in the field of AI liability, seeking to discern prevailing trends regarding the evidence and explanations considered essential for the fair resolution of disputes. Against the backdrop of those trends, the paper evaluates the upcoming legislation in the European Union (EU) concerning AI liability, namely the AI Liability Directive (AILD) and Revised Product Liability Directive (R-PLD). The objective is to ascertain whether the systems of evidence and procedural rights outlined in this legislation, particularly the right to request the disclosure of evidence, enable litigants to adequately understand the causality underlying AI-related harms. Moreover, the paper seeks o determine if litigants can effectively express their views before dispute-resolution authorities based on that understanding. An examination of the AILD and R-PLD reveals that their evidence systems primarily support *ad hoc* explanations, allowing litigants and courts to assess the extent of the defendants' compliance with the standards enshrined in regulatory instruments, such as the AI Act. However, the paper contends that, beyond *ad hoc* explanations, achieving fair resolution in AI liability disputes necessitates *post-hoc* explanations. These should be directed at unveiling the functionalities of AI systems and the rationale behind harmful automated decisions. The paper thus suggests that 'full' explainable AI (XAI) that is, both *ad hoc* and *post hoc*, is necessary so that the constitutional requirements associated with the right to a fair trial (access to courts, equality of arms, contradictory debate) can be effectively met.

Keywords: AI, Causation; Explainability; Fair Trial; Procedural Fairness; Equality of Arms; Effective Participation; AI liability; Product Liability; AI Act; AI Liability Directive; Product Liability Directive

* National Foundation for Scientific Research (FNRS); Faculty of Law, Political Science and Criminology, University of Liège, Belgium.

Table of Contents

INTRODUCTION

A. The Limits of Causal Knowledge and the Refuge of Ignorance Metaphor

In his *Ethics*,[1] 17-century philosopher Spinoza discussed what he termed the 'reduction to ignorance' method, citing an incident of an unfortunate passerby fatally struck by a stone dislodged from a roof. To causally explain the bad timing of the fall, God-fearing dogmatics would, no doubt, ask an endless string of 'why-s': "perhaps you will reply that it happened because the wind blew and the person was walking along that way. But they will press: why did the wind blow at that time? Why was the person going that way at that very time? (…) And so on and so on, and they will not stop asking for causes of causes until you take refuge in the will of God, which is the *refuge of ignorance*."[2]

Our ambition is not to explore the depths of Spinoza's philosophy, but to draw attention to his stance when discussing the construction of knowledge: one would spare oneself from knowing 'proper' if they relied on the *belief* that all worldly occurrences had, as *causa prima*, a metaphysical, omniscient designer of reality. Even pious jusnaturalists like Grotius and Pufendorf hypothesized that if God did not exist (as *the* authority decreeing oughts and ought-nots), Nature would continue to function according to its inherent rationality.[3] Although Spinoza's philosophy is deist - his concept of 'God' coinciding with that of 'Nature' (*Deus sive Natura*)[4] - his work is reflective of the 17-century rationalist rebellion against naïve religiosity, aiming to uncover the dividing line between (true) knowledge and non-knowledge.

---

[1] Baruch Spinoza, *Ethics. Proved in a Geometrical Order*, (ed. by Matthew J. Ksiner, CUP, 2018).

[2] *Id*., at 37 (emphasis added).

[3] Grotius, arguably, pioneered the hypothesis that moral normativity is irrespective of religious affiliation, going counter the Medieval *zeitgeist* according to which, moral normativity was divinely ordained, as opposed to derived from - because inherent to - Man's (rational) nature. Pufendorf later espoused the same view. See, namely, T.J. Hochstrasser, *Natural Law Theories in the Early Enlightenment* (CUP, 2000) at 84: "Pufendorf was entirely correct to identify Grotius and Hobbes as his crucial predecessors, since both had forced their opponents to fight them on new ground of their own choosing: Grotius by insisting that the source of natural law must be located in a principle to which all nations could assent irrespective of religious affiliation; and Hobbes, by his contention that the individual is capable of creating his own moral world from his personal psychological calculations."

[4] Summarizing Spinoza's philosophy is not our point of focus here. May it suffice stressing that he synonymizes God and Nature, asserting that from the infinite attributes of God (Nature), only two are knowable to us: thought and space. All of what is knowable can be understood as a particular expression either of those attributes. On the issue of gaining knowledge of the *essence* of knowable objects, Winch gives an excellent and pedagogical account of Spinoza's epistemology: "Spinoza distinguishes between '*essentia formalis*' and '*essentia objectiva*' (…) the sense of 'objective' doesn't at all lie in a contrast with 'subjective'; it highlights the relation of an idea to its object, to what it asserts or represents to be the case. The 'formal essence' on the other hand is, as it were, the idea as a distinct mental existent, considered in abstraction from its relation to an object." See Peter Winch, *Spinoza on Ethics and Understanding* (CUP, 2020), at 6. Spinoza, much like other philosophers such as Descartes or Kant, tackled the issue of 'knowledge' and 'representation' of reality, the former being traditionally thought to be 'objective' while the latter 'subjective'. The interrelationship between the two, as analyzed in Spinoza's philosophy, will not be further discussed here. However, this is a useful point to keep in mind as we explore the construction of knowledge *tout court* and of causal knowledge because we find, in the backdrop of the relevant theories, the objective/subjective dilemma which has indeed 'tainted' millennia-long traditions of erudite philosophical thought.

Is believing antinomic to knowing? For early-day rationalists, the answer would likely be 'yes.' Modern-day epistemologists are not as quick to dissociate the two, namely because our ability to know is limited. When we are called to causally explain portions of reality that are, to some extent, unknowable to us (e.g. why did a stone mysteriously fall off a roof?), there will invariably come a point where the explanation we give is based, not on 'what *we know to be true'* but on 'what *we believe to be true*.' In many ways, scientific communities today play a role similar to that of religious institutions in Spinoza's time: they nurture normative belief systems that serve as benchmarks for distinguishing valid, trustworthy information from 'false' counterparts. Since science operates largely without relying on faith, the convictions comprising the body of scientific knowledge, including those related to causality, are embraced only when verifiable and verified. Merely asserting claims without substantiation is typically insufficient for justified rational acceptance.

While modern epistemology has eased its skepticism toward beliefs, it has not yet resolved its inner conflict of striving for absolute certainty or truth, alongside the necessity to make internal epistemic compromises in determining what might qualify as *acceptable* knowledge. The pursuit of perfect, permanent, universal, agnostic, and context-independent knowledge alas remains practically unattainable. This - in many ways tragic - realization is at the core of Spinoza's refuge-of-ignorance metaphor: as we endeavor to understand the world causally, we are driven by an ideal (of absolute truth) while being entangled in the constraints of reality (where our capacity to know is limited). The million-dollar question is then: '*how do we decide what is true, if the attainment of perfect knowledge of causation is impossible*?' Probabilists suggested the notion of *necessity*: there comes a point where, by virtue of experience, we detect repetitive, regular associations which we taxonomize as reliable or stable causal phenomena (in the sense of 'X necessarily causes Y'). In their highest expression, these infallible causalities are labelled as (natural) *laws* or normative, 'universal causal regularities.'[5] However, to further our investigation of necessity in connection to causality, it is essential to bring forth one of the 18th century luminaries: David Hume.

B. The Concept of Necessity in Causation

In his Treaties of Human Nature,[6] Hume wrote:

"Probability… must in some respects be founded on the impressions of our memory and senses, and in some respects on our ideas. Were there no mixture of any impression in our probable reasonings, the conclusion wou'd be entirely chimerical: And were there no mixture of ideas, the action of the mind, in observing the relation, wou'd, properly speaking, be sensation, not reasoning.... The only connexion or relation of objects, which can lead us beyond the immediate impressions of our memory and senses, is that of cause and effect. ... The idea of cause and effect is deriv'd from experience, which informs us, that such particular objects, in all past instances, have been constantly conjoined with each other: And as an object similar to one of these is suppos'd to be immediately present in its impression, we thence presume on the existence of one similar to its usual attendant. According to this account of things, ... probability is founded on the presumption of a resemblance betwixt those objects, of which we have had experience, and those of which we have had none; and therefore 't

[5] Max Kistler, *Causation and the Laws of Nature* (Routledge, 2006), at 77.
[6] David Hume, *A Treatise of Human Nature* (ed. by L. A. Selby-Bigge, Clarendon Press, 1888).

is impossible this presumption can arise from probability. The same principle cannot be both cause and effect of another."[7]

An idea that transpires from the cited gloss is Hume's *assumption of uniformity* of Nature. The repetitiveness of observable events (say, rain does not fall when the sky is clear) justifies the associative reasoning Hume referred to. Our past experiences are the cognitive benchmark against which we interpret and explain any new experience. This type of reasoning can be explained by our all-too-human need to somehow make the new familiar. Repetitive events are ultimately what allows us to make causal generalizations which become our *nomological interpretations of reality*:[8] if the weather is cloudy, we may expect rain, snow or nothing at all, but we can be sure not to expect sunshine. The cause/effect link between 'clouds' and 'no sun' enters our arsenal of so-called *background knowledge*, which we mobilize whenever we encounter causal interrelationships we experience as novel.

His brilliance and insight notwithstanding, Hume's Achilles' heel is precisely his assumption that Nature is casually regular. Based on experience, clear skies consistently indicate the absence of rainfall and this we take to be a 'universal given,' a sort of intuitive law by virtue of which rain is generally not expected on a sunny day.

Reality is of course 'messier'[9] than our perceptions thereof, as modern scholarship pointed out in its critique of Hume's work. Kistler e.g. criticized Hume's disregard of *exceptional situations i.e.* cases where real-world occurrences deviate from what we view as nomological causations (*i.e*. causations characterized by a level of predictability).[10] Quantum physics is frequently referenced as an instance of epistemic departure from Newtonian physics: at the sub-atomic level, the behavior of particles appears to deviate from the laws governing supra-atomic behavior.[11] In the context of these 'exceptional situations,' Hume seems to have also omitted *accidental causation i.e.* cause/effect links that we explain in reference to so-called universal laws of Nature. Here again, Kistler cautioned against 'universalizing' the truth of causal phenomena that are due to coincidence[12] and not some unwavering, universal law of Nature.

With Kistler's criticism in mind, it follows that in a perfectly ordained, predictable world, events would, indeed, be causally linked by *necessity*: specific causes would *reliably* yield specific effects and only those. Of course - and again - arriving at a stable universal causal knowledge is a tricky business, for the reasons Kistler outlined in his excellent study.[13]

---

[7] *Id*., at 89-90 *cit. in* Henry W. Johnstone Jr., "Hume's Arguments Concerning Causal Necessity" 16-3 *Philosophy and Phenomenological Research* (1956), 331-340, at 337.

[8] For Kistler, 'nomologcial' is understood as 'normative' within the meaning of the laws of Nature. In the context of causality, we will use 'nomological' to refer to normative representations of necessary cause-effect interrelationships. See Max Kistler, *Causation and the Laws of Nature*, *cit. supra*, at 5.

[9] We paraphrase F.H. Bradley, "Epistemology Legalized: Or, Truth, Justice, and the American Way" *in* Susan Haack, Evidence Matters (CUP, 2014), 27 at 30.

[10] Max Kistler, *Causation and Laws of Nature, cit. supra,* at 76.

[11] For a comprehensive analysis of Newtonian mechanics and Quantum mechanics, see Albrecht Lindner, Dieter Strauch, A Complete Couse on Theoretical Physics. From Clasical Mechanics to Advanced Quantum Statistics (Springer, 2018), at 69 seq. and 275 seq.

[12] Max Kistler, *Causation and Laws of Nature*, *cit. supra*, at 75.

[13] *Id*.

Nevertheless, there is some virtue in epistemic and cognitive stability. Be it in the discovery of causation in science or in law, we cannot consider that all causal relations are a matter of chance. Generalizations about the world (such as clouds usually, though not always mean 'rain') are necessary for our every-day decisions and predictions. Hume's philosophy may be flawed, but it expressed the right intuition: *we need* to consider some causal interrelationships as true. The alternative - a perpetual state of uncertainty and doubt - would simply be untenable. As a result, we choose to assign truth values selectively to specific representations of causality (like 'dark clouds *ergo* rain').

Epistemologists have heavily reflected on the concepts of truth and falsity in causal contexts. Special focus has been placed on the *conditions* under which we decide to designate something as true. This point will be discussed further[14] as we explore the interrelationship between experience, belief and knowledge in explaining causal phenomena. At this stage, we shall stress two points which will frame our further discussion. First, Hume's concept of *causal necessity*, though debatable, has shaped the ways in which we approach knowledge of causation in both 'hard' science and law. Indeed, we often construct such knowledge in terms of necessity (as in 'X necessarily causes Y') because our aim is to ultimately distinguish *correlation* from *causation:* an event can be correlated (positively associated) to several other events but it will be *causally linked* to only one or a few of them. Causes are, in essence, *conditions that appear to be necessary* for specific effects to occur. In law, it is this Humean understanding of 'cause' that underlies the but-for test, which we will discuss further in this paper.[15]

Second, AI poses a challenge to our Humean (and human) inclination to *a priori* perceive reality as relatively stable. To begin with, AI systems exhibit a profound departure from Humean principles, since they are not natural entities, subject to the governance of natural causality. Put differently, we cannot resort to the laws of physics to, say, uncover the origins of algorithmic biases. If AI systems operate outside the jurisdiction of physical laws (as far as causality is concerned) they - intelligent as they are - are, in principle, governed by the laws of (human) reason. In this regard, AI systems align with Humean principles because their decisions and predictions result from associations between existing knowledge (represented by sets of training data) and new information. Just as humans explain new experiences by drawing connections to familiar ones, AI systems create associations between variables in a new situation (unseen during training) and the variable connections already established in the training data. However, the 'laws of reason' do not work as predictably as the laws of Nature, which is inconvenient when we are asked to causally explain the real-world consequences of AI. We thus find ourselves in a conundrum: we are and will increasingly be pushed to causally explain AI 'behavior' without any real possibility of mapping out, if not the 'laws' at least some *consistent trends* regarding the effects that behavior might cause. We know that recruitment AI systems *can be* discriminatory, but they can also be perfectly skill-based…

In dealing with such unpredictability, European and global regulatory reactions were in a manner of speaking, Humean that is, *stability seeking* (as will be argued). They chose to view the uniqueness and novelty of AI rationality through the lens of

[14] See *infra*, Sub-Section 2.2.
[15] See *infra*, Sub-Section 2.2.2.

agency, the referent for stability here being the role of *human* agency in causation. The regulatory verdict was clear: while causal phenomena might involve AI systems, *causal responsibility* will always fall on humans. In light of this, the *causal knowledge* involving AI should allow the identification of a responsible human agent, without it being necessary - or even desirable - to determine if a specific consequence (like harm) was caused by an AI system *having acted alone*. End of the story.

C. AI Output as the Object of Inquiry

Fast-forward a few centuries from Spinoza and Hume: our explanatory abilities have no doubt improved, only nowadays, it is not falling stones but Artificial Intelligence (AI)[16] that pushes us to the edge of what is knowable and explainable. In particular, in the field of AI liability, Spinoza's 'reduction to ignorance' method seems to be far from *dépassé*: just as, centuries ago, divine volition and action were assumed to be the original cause of all worldly occurrences, human intent, action or inaction (in other words, human agency) are now assumed to be the root cause of all harm occasioned by the use of AI systems. Not because we have conclusive evidence that this is always true, but because such is our millennial, *normative belief*: people harm people, even if the causing of harm is made possible through the use of sophisticated, smart technologies.

Our collective preference to uphold an anthropocentric view of causality is perhaps a 'healthy' reaction to the realization that AI systems can work in mysterious ways. Examples of recruitment AI, automated vehicles and credit-scoring AI, to name a few have shown that intelligent systems may not always offer the possibility for their decisional processes to be scrutinized. To compensate our lack of causal knowledge in such instances, we turn to our 'nomic' causal representations, *seeking refuge in the human agency postulate*, as cornerstone of longstanding liability doctrines.[17] But those doctrines date from a time when non-human intelligence and agency were inconceivable… In recent decades, part of scholarship reflected on whether the concept of agency ought to be reconceptualized in order to extend to non-human entities who reason (and therefore, act) in similar ways as humans. The consensus has fallen on the

[16] For the purpose of this paper, we will refer to the definition of AI included in the AI Act. See, Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (AI Act), and amending certain Union legislative acts, COM(2021) 206 *final*, art. 3(1): "artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." The 'techniques and approaches' mentioned in Annex I are Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning (Annex I, a)); logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems (Annex I, b)) and statistical approaches, Bayesian estimation, search and optimization methods (Annex I, c)).

[17] See *inter alia* Ljupcho Grozdanovski, «L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale? Réflexions sur l'avenir du régime de responsabilité du fait de produits défectueux dans l'Union européenne» (2022) 232/233 2 *Réseaux,* 99.

fact that, their levels of intelligence[18] notwithstanding, AI systems form a class of commodities[19] meaning that, when harm is causally associated with those systems, the culprit will invariably be the human having either programmed or used them. But in doing so, are we not choosing a *causal belief* over *causal knowledge*? Are we not (re)creating a 'refuge of ignorance'?... It certainly seems so. In lieu of looking to design discovery methods through which litigants could uncover the *actual causal power* of AI systems we, as a collective, seem to prefer the safety of what we have always known to be true *i.e*. that rational and moral agency can only be a human prerogative.

The postulate of the 'human puppeteer' - discrete but always present behind the scenes in opaque AI decision-making - could perhaps be tenable, had we remained in the early days of Artificial Narrow Intelligence (ANI). In the stone age of AI - dating to only a few years ago - we mostly dealt with hyperspecialized "idiot savants,"[20] very good in performing one task or a set of tasks, useless at anything else. Since then, technological innovation has developed at a galloping pace, resulting in more generally intelligent systems. Generative AI like ChatGPT gives an illustration of this. We have not yet reached the stage of Artificial General Intelligence (AGI)[21] and certainly not that of Artificial Super Intelligence (ASI)… But we are getting there. Of course, 'general intelligence' is a multifaceted concept which includes - under the 'general' label - several types of intelligence.[22] For the sake of simplicity, we will consider the level(s) of AI intelligence as correlating to level(s) of *cognitive* and *decisional*

---

[18] Though there are many possible ways to define intelligence *tout court*, it is possible to argue that it translates to a series of abilities that allow an agent to *autonomously arrive* at a solution or make a prediction in a context where all the variables are not known. See Kristin Thorisson, Helgi Helgasson, "Cognitive Architectures and Autonomy: A Comparative Review" (2012), 3-2 *J. Gen. AI*, 1, at 3. Intelligence in connection to (artificial) agency has raised issues on whether AI's autonomy can warrant the recognition of some form of agency. We have argued in our previous work that AI's autonomy is similar to human autonomy *functionally* in that AI systems are able to simulate human skills which, when exercised - and as a general rule of thumb - aim for efficiency and accuracy. AI's ability to replicate human intelligence has not yet extended to *human ontology*, placing in the core of what it means to be 'intelligent' the (autonomous) ability for empathy and more generally, the ability to distinguish right from wrong. On the distinction between functional and ontological aspects of human and non-human intelligence, see Ljupcho Grozdanovski, «L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale? Réflexions sur l'avenir du régime de responsabilité du fait de produits défectueux dans l'Union européenne», *cit. supra*, at 9.

[19] Commoditization of advanced technologies is not recent. One of its oldest expressions can be found in American caselaw which interpreted robots as mechanical devices "a mere automation, that operates through scientific or mechanical media" but is not "a living thing; it is not endowed with life." See *Louis Marx & Co. and Gehrig Hoban & Co., Inc. v. United States* case (40 Cust. Ct. 610, 610 (1958)). For a comment on this and other US cases in the field of robotics, See Ryan Calo, "Robots in American Law," *Legal Studies Research Paper N° 2016-4* (University of Washington - School of Law), available on: http://euro.ecom.cmu.edu/program/law/08-732/AI/Calo.pdf (last accessed on 20 Jan. 2024), at 14.

[20] Matt Paisner, Michael T. Cox, Michael Maynord, Don Perlis "Goal-driven autonomy for cognitive systems", Proceedings of the Cognitive Science Society (2014), available at <pdfs.semanticscholar. org/2c9c/2bb5381a0e094d80b2095dbedbbe6546911e.pdf>, 2085–2090, at 2085.

[21] AGI includes AI systems able to perform most, if not all, cognitive functions as good as humans, Gonenv Gurkaynak, Ilay Yimaz, Gunes Haksever "Stifling Artificial Intelli- gence: Human Perils" (2016) 32-5 *Comp. L. & Sec'y Rev.*, 749, at 751.

[22] The taxonomy of intelligence is a delicate issue, in the sense that clear-cut categories or types of intelligence are difficult to establish. There are, however, several types of 'abilities' which scholars have associated with types of intelligence. They include, namely, so-called fluid intelligence, crystallized intelligence, visual intelligence, auditory intelligence, cognitive processing speed etc. See Wan Nurul Izza Wan Husin, Angeli Santos, Hazel Melanie Ramos, Mohamad Sahari Nordin, "The place of emotional intelligence in the 'intelligence' taxonomy: Crystallized intelligence or fluid intelligence" (2013) 97 *Procedia - Soc. & Behav'l Sci.*, 214, at 215.

autonomy in reaching a preassigned goal and, in some cases - like those of Deep Learning (DL) systems[23]- even selecting the goal(s) to be achieved. As we will argue further, the more generally intelligent the system, the greater its level of autonomy and the more accurate its outcomes but also, the less scrutable the reasoning patterns through which those outcomes are arrived at.

In sum, we seem to be caught in a thug of war between, on the one hand, imminent technological evolution which promises to emancipate AI form any realistic form of 'panoptic' human control and oversight and, on the other hand, a regulatory *penchant* for stability and continuity, characterized by AI commoditization and the sacrosanct human agency principle. This, of course, has an important impact on the *design of the systems of evidence* used in the adjudication of disputes dealing with AI liability.

D. The Possibility for Evidence and (Causal) Explanation Pertaining to AI Output

The concept of legal evidence[24] is a curious beast, because it simultaneously answers to two sets of validity criteria: those of truth and those of fairness. The realm of truth is that of discovery and epistemology[25] which, in the field of procedural law, find a specific expression in legal rules and principles of evidence. The *raison d'être* of those rules and principles is to *epistemically frame* the process of fact-finding and fact-assessment under an *independent* (impartial) standard of accuracy. Of course, in adjudicatory contexts, fact-accuracy is not sought for accuracy's sake: 'accurate' knowledge of the disputed facts is a factor that impacts the fairness of a dispute's *outcome*.[26] This accuracy/fairness interplay is precisely what marks the specificity of legal evidence as a concept: fairness is both the expected *outcome* from an institutional - most commonly, judicial - law-to-fact application and the *epistemic constraint* of the process through which knowledge of the disputed facts is construed. The longstanding normative creed is, indeed, that only fair procedures (*i.e.* designed to create conditions of fair adjudication) can be conducive to fair outcomes.[27]

Concretely, this means that the parties in a dispute should have *equal procedural abilities* to access and give the evidence they view as relevant and probative. This

---

[23] DL systems are models with multilayered neural networks that are trained with large data sets of data and able to solve highly complex information processing tasks. For an analysis of DL models in fields like medicine, see Christopher M. Bishop, Hugh Bishop, *Deep Learning. Foundations and Concepts* (Springer, 2024).

[24] According to Wigmore, 'evidence' can be understood as any knowable fact or group of facts, considered with a view to its being offered for the purpose of producing conviction as to the truth of a proposition. See John Henry Wigmore, *Evidence in Trials at Common Law* (Little, Brown, 4th ed., 1961).

[25] Epistemology will be understood as the field of study focused on the theorizing and structuring methods of knowledge and beliefs construction. See, *inter alia*, Jaakko Hintikka, *Socratic Epistemology. Explorations of Knowledge-Seeking by Questioning* (CUP, 2012) at 11 seq.

[26] In some strands of evidence scholarship, accurate representations of fact are needed to give way to a correct application of the law, the belief here being that - as Grando put it - "accurate decisions are usually fair." See Michelle T. Grando, *Evidence, Proof, and Fact-Finding in WTO Dispute Settlement*, (OUP, 2009) at 11.

[27] The fair procedures/fair outcomes parallelism derives from Rawls' idea(l) of so-called *perfect procedural justice* model by virtue of which fair procedures, if correctly followed, yield correct and fair results. See John Rawls, *A Theory of Justice* (revised ed.) (Harv. UP, 1999), at 75 seq.

procedural parity, typically expressed in the fair-trial safeguards,[28] is meant to define a level of *baseline equality*, placing the parties on an equal procedural footing when they make their views known before an adjudicating authority. From the evidentiary debate thus organized - and conceptually akin to Habermas's discursive ethics[29] - 'truth' is expected to surface, giving courts the information necessary to answer two cardinal questions: '*what and who caused the dispute*?' and '*what is the most adequate (or fair) legal solution to that dispute*?'

With the truth/fairness interplay in the backdrop, let us turn to the *evidence of causation*. Two issues can be flagged as relevant. First, there is the already discussed (Humean) issue of *necessity*, which invites us to reflect on the evidence and corresponding explanations litigants should *be able to access* to effectively argue how an event was causally linked to another event (typically, a harm). Second, there is - again - the issue of *fairness*: how should systems of evidence, in the EU, be (re)designed so that the *evidence flagged as necessary* under point (1) can be adduced in conditions of procedural parity? To answer both questions - as this paper's chief ambition - we must address a more fundamental issue, characteristic of AI liability: *what exactly are we seeking to explain when we give evidence on the casual link between an AI system and a harm*? Two roads diverge[30] here: the one, more travelled, asks us to explain causality from the vantage point of human agency; the other, less travelled, asks us to engage in proper discovery of the causal chain between a harm and a harm-causing conduct (possibly of a non-human, intelligent entity).

We already alluded to the first alternative earlier: in lieu of engaging in Byzantine debates on whether harm can be imputable to an AI having acted alone, we seem to prefer the *belief* that the authorship of (and by that, the responsibility for) that harm is incumbent to a human (programmer, user), without this warranting an in-depth demonstration of whether that human's actions *actually* contributed to the harm-causing automated decision. Taking human agency as a presumed (as opposed to established) cause of such harm is, of course, reassuring because it maintains conceptual continuity, but it barely holds in the scenario where there is *no evidence of human involvement*, and yet harm was somehow occasioned by an AI's use.

The second alternative is the one where the evidentiary debate on causality would include discovery proper, yielding explanations on *how* a given system made a harmful decision or prediction. Part of AI scholarship supports this view. For example, Barredo Arrieta *et al*. made a point on the nature of causal knowledge, by making the distinction between *causality* and *causation*. Causality, the authors argue, requires a

---

[28] In the EU, the fair trial safeguards are currently enshrined in Article 47 of the EU Charter of Fundamental rights (EUCFR). Those safeguards include the right to a fair and public hearing within a reasonable time by an independent and impartial tribunal previously established by law (Art. 47(2) EUCFR).

[29] We refer to Habermas' 'ideal speech situation' based on three (participatory) equality-enhancing rules namely, the rule of participation, the rule of equal opportunity and the rule against compulsion. See Jürgen Habermas, *The Theory of Communicative Action: Reason and Rationalization of Society*, Beacon Press (1984).

[30] This is an expression drawn from Robert Frost, "The Road Not Taken" (2021) 4 *The Objective Standard*, at 79.

"wide frame of prior knowledge to prove that observed effects are causal."[31] Causation "involves correlation, so an explainable ML model could validate the results provided by causality inference techniques, or provide a first intuition of possible causal relationships within the available data."[32]

When a court seeks to determine 'what happened' in an AI liability case, the knowledge that they would normally seek is that of causation, as defined by Barredo Arrieta *et al*. The practical problem here is that the discovery of causation may not be feasible because the evidence thereof may not be - reasonably - within the litigants' reach. As mentioned earlier, the variable-correlations an AI system may have made prior to the occurrence of harm often remain partially or fully unknowable to human agents (sometimes, including the programmers). For example, how could a loan applicant *even suspect* that an AI, used to preapprove loan applications, was racially biased? That applicant would presumably have no access to the applicants the system had approved, nor would they have information of how the bank usually assesses applicants' credit. In this context, to make their argument, the claimant would require access to two types of evidence. First, they would need to establish that the system's output was indeed racially biased, which implies that they should, somehow, understand and explain that the outcome of a specific variable association (e.g. place of residence *cum* ability to repay the loan) was a key factor in the occurrence of racial discrimination. Second, to causally explain that discrimination, they would need to establish and explain *what actually caused* it (*i.e.* explain if the bias was embedded in the programming data or machine learnt.). If there is evidence showing that the bias was machine learnt, who should then be held as liable?...

We have examined the issue of allocating liability elsewhere.[33] Our suggestion was that, when the human authorship of AI-related harm is not proven, the liable agent (*i.e.* held to compensate the harm) should be the *one having accepted the risk* of the harm occurring. That agent can be either the *programmer*, having released in the market a system that has, in the past, been prone to certain types of malfunctions (e.g. developing unfair biases) or the *user* who, aware of the harms a system may typically cause, had chosen to nevertheless use it.

In this paper, our focus will be more on the *evidentiary causal explanations* needed to determine the *locus* of AI liability, under the European Union's (EU) regulatory framework. In this context, we will explore what can and should be established and explained, when the chain of causality is fully or partially unknowable - possibly more so than in 'ordinary' causal scenarios (*i.e*. those that do not include intelligent systems).

---

[31] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, *et al*. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (2020) 58 *Information Fusion*, 82, at 86.

[32] *Ibid*.

[33] Ljupcho Grozdanovski, «L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale? Réflexions sur l'avenir du régime de responsabilité de produits défectueux dans l'Union européenne» *cit. supra*.

E. A Shift in Perspective: From Causal Explanations Required by Law to Causal Explanations Asked for (and Given) by Litigants

Bearing in mind our 'two roads diverge' metaphor, the legislator of the EU - much like the legislators of several countries around the world - was faced with the difficult task of regulating AI liability against the backdrop of two competing principles: that of discovery (causal knowledge) and that of human agency (belief). In the field of procedure, the guiding principle in choosing the one over the other should, no doubt, be that of (procedural) *fairness*.

If we draw on standard liability doctrines and consider that fair outcomes *always* call for accurate knowledge of causation, then AI liability should not be viewed as an exception, meaning that the culprit should be identified through evidence, not presumptions. If, however, a standard of fairness is thought to be best upheld when the law's postulates remain unshaken, then the level of causal accuracy in AI liability will be required *to the extent that it coheres* with the presumption of human agency… But this is AI liability viewed from the heights of the conceptual tower that is (standard) liability law. It is perhaps more relevant to inspect what happens in the trenches *i.e.* in the *already adjudicated* and/or *forthcoming* AI liability disputes. These invite us to set aside the deontic stance of the law and take on a more down-to-earth, fact-based and, dare we say, humanist perspective by addressing the oft-forgotten '*what do the people need*?' question. Do the *litigants themselves* consider that, to argue causation, they need to understand *how* an AI system caused harm or is this knowledge procedurally irrelevant to them?...

Fundamentally, this paper seeks to conceptualize *procedural fairness* in the face of AI and to do so, it will follow a *bottom-up approach*. It will depart from court practice - mainly North-American - and will seek to induce the features of a concept of 'AI fairness' based on the *procedural needs* expressed by litigants in AI liability cases. Against this backdrop, this paper will critically assess the EU's AI liability regulatory framework, sketching out ways in which that framework ought to be applied, in view of better supporting the litigants' so-called *effective participation*[34] in the resolution of future AI liability disputes.

The doctrinal strand that we will take as a key analytical referent is the doctrine of so-called *procedural abilities* - basic entitlements litigants ought to have to effectively make their views known before a court. This school of thought developed as the procedural 'spinoff' of the so-called *capabilities* approach, as conceptualized in the seminal work of Sen[35] and Nussbaum.[36] Unlike previous - say, Rawlsian[37] - justice theories, aimed at distilling normative, universal understandings of fundamental principles of justice like 'the right,' 'the equal' and 'the good,' the capabilities strand is more interested in the *entitlements* individuals should enjoy to live 'meaningful' lives, the real-world injustices notwithstanding. In a taxonomical *élan*, Nussbaum seminally suggested ten fundamental capabilities which, she argued, are the universal prerequisites for a thriving human existence. These are: life, bodily health, bodily

---

34 Lawrence Solum, "Procedural Justice" (2004) 78 *Calif. L. Rev.*, 181, at 305.
35 Amartya Sen *The Idea of Justice* (Harv.U.P, 2009).
36 Martha C. Nussbaum, *Frontiers of Justice. Disability, Nationality, Species Membership* (Harv.U.P., 2006).
37 John Rawls, *A Theory of Justice*, *cit. supra*.

integrity, senses, imagination and thought, emotions, practical reason, affiliation, play and control over one's environment. The capabilities approach has also been the object of criticism. However, one of its merits is that it offers, if not a perfect, at least a *workable understanding of fairness*, acting more as a general guideline for regulatory action, than a mandatory ethical precept. This is no doubt the reason why Sen's and Nussbaum's scholarship laid the theoretical foundation for the United Nations' (UN) Sustainable Development Goals (SDGs).

In the field of procedure, the capabilities approach was echoed in the so-called procedural abilities - basic procedural entitlements that parties in adjudicatory contexts should have to 'meaningfully'[38] participate in adjudicatory processes. Mirroring Nussbaum's decalogue, Awusu-Bempah[39] suggested a taxonomy of procedural abilities which are also ten: 1) understand the nature of the charge; 2) understand the evidence adduced; 3) understand the trial process and the consequences of being convicted; 4) give instructions to a legal representative; 5) make a decision about whether to plead guilty or not guilty; 6) make a decision about whether to give evidence; 7) make other decisions that might need to be made by the defendant in connection with the trial; 8) follow the proceedings in court on the offence; 9) give evidence; 10) any other ability that appears to the court to be relevant in the particular case.[40] The choice of the procedural abilities strand as the 'intellectual compass' of our analysis is justified by our preoccupation with *effectiveness* translated in, what we previously labelled as, our bottom-up approach to conceptualizing AI (procedural) fairness.

As a matter of *personal conviction* of this paper's author: litigants should feel that the law gives them a discursive space where they can speak their truth.

As a matter of *factual accuracy of AI causation*: litigants should feel that important decisions like those on responsibility or guilt are not arbitrary but informed, based on accurate information.

As a matter of *procedural fairness* in the face of AI: litigants should feel that a system of procedures and remedies provides them with the abilities *they need* to discuss matters like innocence and guilt.

In this context, rather than investigating how (procedural) law should align itself concerning the proof of causality or the presumption of human responsibility, it may be more prudent to contemplate what litigants engaged in discussions about AI-related harm *should be capable of proving and explaining* to ensure a fair resolution to their dispute. This shift from the 'procedural ought' to the 'procedural need' naturally pushes us to raise the issue of the *access* to evidence: if an AI's inner workings are unknowable, how can non-expert litigants access the information *they need* to provide an explanation on who or what caused the harm? Should law provide a procedural right to access such evidence?... These and other questions will be raised in our analysis of the interrelationship between 'what is' explanation in connection to AI, and 'how' that explanation ties (or not) into causal explanations of harm given in AI liability disputes.

---

38 Lawrence Solum, "Procedural Justice," *cit. supra*, at 305.

39 Abenaa Owusu-Bempah, "The interpretation and application of the right to effective participation" (2018) 22-4 *The Int'l J. of Evidence & Proof*, 321.

40 *Id*., at 330.

However, before we outline the structure of our arguments on this point, it is necessary to say a few words on the EU's regulatory frameworks of AI.

F. The EU's Regulation of AI

1. The Substantive Regulation - the AI Act

AI regulation in the Union essentially evolved in two stages. First came substantive law in the form of a proposal for a Regulation laying down harmonized rules on AI (AI Act).[41] We have extensively explored the history and content of this instrument elsewhere and will not offer a detailed account thereof here. We will but mention the aspects of the AI Act that we view as relevant for the remained of this paper.

On the *type of regulation*, the AI Act can, in essence, be thought of as an instrument that transposes product safety logic to risks of fundamental rights violations. The operative assumption is that, like 'ordinary' products, AI systems can be safely used if their programing and use comply with a number of predefined technical standards. This of course is debatable, but we will refrain from further commenting on whether the tried-and-true method of standardized product manufacturing is a good fit for regulating products which are not automated but intelligent. This was *inter alia* a point raised in one of our recent studies.[42]

---

41 Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final.

42 Ljupcho Grozdanovski, Jérôme de Cooman, "Forget the Facts, Aim for the Rights! On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union" (2023) 2 *Rutgers Comp. & Tech'y L. J.*, 207.

More importantly, the AI Act includes a four-level taxonomy of risks: non-high,[43] limited,[44] high and unacceptable.[45] The so-called high-risk AI systems are the most relevant for this paper because the evidentiary frameworks included in the EU's procedural regulation following the AI Act were specifically designed to enable proof of causation in cases involving those systems.

High-risk AI is a class of intelligent systems assumed to pose threats of fundamental rights violations and yet their commercialization is allowed: "rather than being altogether prohibited, they are subject to mandatory requirements, chiefly transparency (art. 13) and human oversight (art. 14)."[46] The AI Act - we argued in our study - distinguishes between two categories of high-risk AI: "the first category includes systems intended to be used as safety component of products covered by EU sectorial product legislations listed in Annex H (art. 6(1)(a)) and that are subject to third party *ex-ante* conformity assessment (art. 6(1)(b)), bearing in mind that a safety component is 'a component of a product or of a system which fulfils a safety function for that product or system or the failure or malfunctioning of which endangers the health and safety of persons or property" (art. 3(14))."[47] The second category "includes stand-alone Al systems with mainly fundamental rights implications that are explicitly listed in Annex III (art. 6(2))."[48]

Annex III of the AI Act lists *eight key areas* where high-risk systems are most likely to be used: biometric identification and categorization of natural persons; management and operation of critical infrastructure; education and vocational training; employment, workers management and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law

---

43 *Id*., at 243: "*non-high-risk AI systems* are defined in opposition to high-risk systems. As high-risks Al systems are exhaustively enumerated, non-high-risks Al systems form a residual (and presumably the largest) category. The regulatory principle for those systems is the absence ofa duty to comply with the mandatory requirements which target the high-risks systems (Art. 8). Developers and users of non-high risk AI systems are, however, encouraged to voluntarily apply these requirements through codes of conduct (Art. 69)."

44 *Id.* at 243-244: "*Limited risks AI system* are, similarly, not subject to mandatory requirements set up in the AI Act (art. 8). However, the Al Act does establish an obligation of transparency for systems which, though formally qualified as non-high risk, interact with natural persons (art. 52(1)), perform emotion recognition or biometric categorization (art. 52(2)). Such systems ought to be designed in a way that natural persons know they interact with or are exposed to an Al system. In a similar vein, users of so-called deepfake technology - *i.e.,* hyper- realistic videos using face swaps that leave little trace of manipulation - are required to disclose that the content has been manipulated or artificially generated (art. *52*(3))."

45 *Id*., at 244: "Al systems that pose *unacceptable risks* are subject to an *ex officio* ban (art. 5). It should be stressed that military applications are excluded from the scope of the Al Act (art. 2(3)). With this exception in mind, AI systems that either use subliminal manipulation of natural person's consciousness (art. 5(1)(a))) or exploit vulnerabilities of a specific group of persons due to their characteristics, *e.g.,* age, physical or psychological disability (art. 5(1)(b)) in order to distort people's behavior in a way that is likely to cause physical or psychological harm are prohibited. The ban also extends to AI systems used by public authorities that score natural persons based on their personal and social behavior, known or predicted (art. 5(1)(c)) as well as those that may lead to detrimental or unfavorable treatment of certain natural persons or groups either "in social contexts which are unrelated to the contexts in which the data was originally generated or collected" (art. 5(1)(c)(i)) or that is "unjustified or disproportionate to their social behavior or its gravity" (art. 5(1)(c)(ii))."

46 *Id*., at 244.

47 *Ibid*.

48 *Ibid*.

enforcement; predictive policing and migration, asylum and border control management. For the systems used in these sectors, the AI Act defines technical standards for compliance such as risk-management (Art. 9), data and data governance (Art. 10), technical documentation (Art. 11), record-keeping (Art. 12), transparency and provision of information to users (Art. 13), human oversight (Art. 14), accuracy, robustness and cybersecurity (Art. 15).

The European Commission's initial proposal for the AI Act underwent several modifications from the EU's legislative bodies *i.e*. the Parliament and the Council. A provisional agreement was eventually reached on 9 December 2023.[49] However, as of that date, a definitive consolidated version of the AI Act was not released; only a document compiling the specific agreed-upon amendments was disclosed. On 22 January 2024, an unofficial version of the AI Act was leaked by EurActive editor Luca Bertuzzi.[50] For the remainder of this paper, we will refer to this leaked version when citing specific provisions from the AI Act.

By identifying the sectors where the risk of AI-related harm is 'high,' the AI Act is, without a doubt, a laudable first since it transcends the congenital diversity of AI as a class of new technologies. However, this instrument relies on a somewhat fallacious assumption: that compliance will somehow suffice for harm to be prevented. Because of this, the AI Act contains virtually no provisions on the *ex post* protection of human agents when a harm eventually ends up materializing. To fill this gap, in September 2022, the EC published a Directive Proposal on adapting non-contractual civil liability rules to AI (AI Liability Directive - AILD).[51]

2. The Procedural Regulation

a. The AI Liability Directive - AILD

The AILD echoes the regulatory principles enshrined in the AI Act and, much like this instrument, it seeks to strike a balance between increasing market gains (by fostering competitiveness and investment in research and innovation), and the safeguard of - what we may call - *non-waivable fundamental rights and democratic values*. In this context, the AILD explicitly states that "to reap the economic and societal benefits of AI and promote the transition to the digital economy, it is necessary to *adapt in a targeted manner certain national civil liability rules to those specific characteristics of certain AI systems*."[52] According to this Directive, the point of reconciliation between market efficiency and procedural fairness is *trust*.[53] The 'adaptations' of national civil liability rules the AILD mentions are assumed to

---

[49] See https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf (last accessed on 23 Jan. 2024).

[50] See Jedidiah Bracy, "EU AI Act: Draft consolidated text leaked online," available on: https://iapp.org/news/a/eu-ai-act-draft-consolidated-text-leaked-online/ (last accessed on 23 Jan. 2024).

[51] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence (AI Liability Directive - AILD) COM (2022) 496 final.

[52] *Id*., Preamble, pt 5 (emphasis added).

[53] This echoes the key objectives highlighted to frame the EU's Regulation of AI, namely - what the EC called - the ecosystem of trust and the ecosystem of excellence. See European Commission, White Paper, *On Artificial Intelligence - A European Approach to Excellence and Trust*, COM (2020) 65 final.

contribute to "*societal and consumer trust* and promote the roll-out of AI"[54] but they are also assumed to "*maintain trust in the judicial system*, by ensuring that victims of damage caused with the involvement of AI have the same effective compensation as victims of damage caused by other technologies."[55] A 'workable equilibrium' between these two 'pillars of trust' can - the Directive states - be achieved through the harmonization of certain non-contractual *fault-based* liability rules, aimed at ensuring that persons who claim compensation for harm caused by AI systems "enjoy a *level of protection equivalent to that enjoyed by persons claiming compensation for damage caused without the involvement of an AI system*."[56]

The AILD carries the imprint of the initial regulatory impulse given by the early-day EU instruments on AI (namely the HLEG's Guidelines on Ethics[57] and the White Paper on AI[58]): the achievement of market gains is not pursued in parallel to a 'pedagogical' protection of fundamental rights; rather *the realization of market gains is framed by the protection of those rights*. This is a relevant point of comparison with the AI Act which, following a logic of prevention of AI-related risks, defines the notion of 'risk' precisely as a violation of fundamental rights and values.[59] That understanding of risk has largely shaped the design of the system of evidence contained in the AILD.

Two main features of this Directive will be highlighted, at this stage. First, it creates a *fault-based* - as opposed to strict - liability regime. This means that the compensation of harm occasioned by an AI system will require *proof of fault*. In this regard, the link between the AI Act and said Directive is salient, given that the notion (and therefore, evidence) of fault is defined as a behavior consisting in "the non-compliance with a duty of care laid down in Union or national law directly intended to protect against the damage that occurred" (Art. 4(2)). The notion of 'fault' is therefore not defined as one might typically expect *i.e.* as the result from a *wrongful* act (*i.e*. a violation of a duty of care, regardless of whether that duty is recognized in a legal provision).[60] Rather, 'fault' is a failure to comply with the standards explicitly laid down in the AI Act's provisions.[61] Fault is therefore understood as *unlawful conduct* (non-compliance with the law) which, as we will subsequently argue, has an important impact on the types of evidence that litigants are authorized to ask for and adduce on the grounds of the instrument considered. Under certain conditions - also discussed further - it is the proof of this type of 'fault' that provides the grounds for a *presumption of causality*.

The justification for this (overly legalistic?) understanding of fault is the fact that the AI Act creates *full harmonization* of the technical requirements pertaining to

---

54 AILD (COM (2022) 496 final) *cit. supra*, Preamble, pt 5.
55 *Ibid* (emphasis added).
56 *Id*., Preamble, pt 7.
57 High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI* (2019), available on: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed on 20 Jan. 2024)
58 European Commission, White Paper, On Artificial Intelligence - A European Approach to Excellence and Trust.
59 Article 2 AI Act, *cit. supra*.
60 See our discussion on wrongfulness and unlawfulness *infra*, Sub-Section 2.2.1 (A).
61 AILD (COM (2022) 496 final) *cit. supra*, Preamble, pt 26: "This Directive covers the fault constituting non-compliance with certain listed requirements laid down in Chapters 2 and 3 of [the AI Act] for providers and users of high-risk AI systems."

the programming and use of high-risk AI systems.[62] Against this backdrop, "and in *full consistency with the logic* of the AI Act,"[63] one of the ambitions stressed in the AILD is to provide steps for providers to adopt or not risk management measures as *relevant evidence* for the purpose of determining whether there has been a case of non-compliance.[64] Further in this paper, we will be critical of the notion of fault, as defined in the AILD and the system of evidence designed around it. At this stage, may it suffice stressing that fault is the point where the AI Act and the AILD intersect: to prove fault under the latter, one would need to prove non-compliance with the former.

Second, the AILD introduces *minimal harmonization* which means that national courts will apply their national rules of procedure and evidence in areas not covered by this harmonization. However, the Directive provides some important procedural guidelines. Two of its key advancements are the *right to request disclosure of evidence* (Art. 3) for victims, and the *right to rebut the so-called presumption of causality*, for respondents (AI providers or users). These are arguably the main source of value of the instrument under consideration. By recognizing the right to request disclosure of evidence, the latter gives a procedural expression to the twin principles of transparency and explainability: after all, only a transparent automated decision can 'open' the access to facts, thus providing grounds for plausible arguments of fault and causation to be presented before a court. However, our analysis of these rights will reveal several inconsistencies in the way the right to request disclosure of evidence is exercised.

Regarding the allocation of the burdens of proof, the AILD defines those - albeit in general terms - by canvassing the main requirements that claimants should meet when arguing and proving fault and causation.[65] The types of relevant facts (*facti probandi*) vary, depending on whether the respondent is an AI provider or an AI user. When the respondent is a provider, the claimant is held to prove the latter's failure to comply with the requirements, listed in the AI Act, that target the so-called 'high-risk' systems. These requirements include transparency (Article 13 AI Act); effective oversight (Article 14 AI Act); accuracy, robustness and cybersecurity (Articles 15 and 16 AI Act); the taking of necessary corrective actions (Articles 16 and 21 AI Act). Alternatively, when the respondent is a user, the claimant is held to prove the former's failure to comply with instructions of use (Article 29 AI Act) and/or exposure of the system to input data which is not relevant in view of the system's intended purpose (Article 29(3), AI Act).

---

62 *Ibid*.
63 *Ibid*.
64 *Ibid*.
65 It should be stressed that the Member States' courts are not deprived of their discretion in defining the relevant facts. However, this discretion notwithstanding, said Directive provides guidelines on the issue of relevance, as regards the proof of fault. Art. 3(1), AILD, "Member States shall ensure that national courts are empowered (…) to disclose relevant evidence (…) about specific high-risk AI systems that is suspected of having caused damage, but was refused, or a claimant, to order the disclosure of such evidence from those persons."

b. The Revised Product Liability Directive - R-PLD

Dating back to 1985, the Product Liability Directive (PLD)[66] naturally did not have the foresight of including AI in its scope of application.'[67] According to the European Commission (EC), the PLD's shortcomings warranting revision mainly had to do with the *design* of the system of evidence the Directive created. In particular, the proof of defectiveness and its link to a harm had shown to be challenging for claimants, especially in complex cases like those involving pharmaceuticals, smart products or AI-enabled products.[68]

Unlike the AILD - which creates a fault-based liability system of evidence - the PLD establishes a *strict liability* system, not requiring proof of fault. The relevant fact (*factum probandum* or the fact for which evidence is sought) that litigants are called to establish within the PLD is *defectiveness*. The proposal for a revision of the PLD (R-PLD) did not change this aspect of the original PLD. The 'new' product liability framework also integrates the strict liability logic, stating that, when AI systems are defective and cause physical harm, property damage or data loss "it is possible to seek compensation from the AI-system provider or from any manufacturer that integrates an AI system into another product."[69]

Prior to submitting the R-PLD proposal, the EC launched a public consultation, during which 77% of participants underlined the procedural challenges faced by litigants in cases involving technically complex products.[70] Pushed to revisit the system of evidence from 1985 - in view of lightening the burden of proof for victims - the EC considered several lines of revision, but ultimately decided on two. First, regarding the *types of products* included in the 'new' Directive's scope of application, the EC chose to include, in the 'new' Directive's scope of application, manufacturers and providers of intangible digital elements, as well as 3^{d} parties providing software added to a product. Second, regarding more specifically the *design* of the system of evidence centered around defectiveness, the EC opted for a system that would ease the burden on consumers by harmonizing the rules on the disclosure of technical information to the victims and the conditions under which defectiveness can be presumed.[71]

To achieve the ambition of 'lightening' the burden of proof especially for

---

66 Council Directive 85/374, of 25 July 1985, on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, OJ L 210, 7.8.1985, p. 29.

67 In Art. 2 PLD, 'product' is defined as "all movables, with the exception of primary agricultural products and game, even though incorporated into another movable or into an immovable (…) 'Product' includes electricity." In the proposal for revision of the PLD (R-PLD), the 1985 definition is broadened. Art. 4(1) R-PLD, states that 'product' means "all movables, even if integrated into another movable or into an immovable. 'Product' includes electricity, digital manufacturing and software."

68 EC, Proposal for a Directive of the European Parliament and of the Council on liability of defective products, COM (2022) 495 final, at 1.

69 *Id*., at 3.

70 *Id*., at 8. The percentage was considerably higher among consumer organisations, NGOs and members of the public (95%) than among business and industry organisations (38%). Industry stakeholders were more open to information disclosure obligations and easing the burden of proof in complex cases than to reversing the burden of proof, which they considered a radical option that would harm innovation.

71 *Id*., pt 9.

claimants, the R-PLD sought to mend the asymmetry of information between the parties in cases characterized by technical or scientific complexity.[72] To do so, it used a well-known procedural 'trick': presumptions. Rebuttable presumptions of fact - the R-PLD states - are "a common mechanism for alleviating a claimant's evidential difficulties, and *allow a court to base the existence of defectiveness or causal link on the presence of another fact that has been proven*, while preserving the rights of the defendant."[73] Indeed, national courts can presume the defectiveness, causation, or both where, "notwithstanding the defendant's disclosure of information, it would be *excessively difficult* for the claimant, in light of the technical or scientific complexity of the case, to prove its defectiveness or the causal link, or both. In such cases, requiring proof would undermine the effectiveness of the right to compensation."[74]

Though the AILD and the R-PLD differ regarding their *facti probandi* (respectively, fault and defectiveness), they converge in two important ways. First, both instruments recognize a right to request a disclosure of evidence (in an *élan* to make evidence more feasible) for claimants. Second, in both instruments, the defendants' refusal to disclose 'relevant evidence' - whatever that might be - generates presumptions (of fault, of defectiveness and/or of causation).

Following up on the 'lightening the burden' idea, Article 9 R-PLD establishes the basic tenets of the upcoming evidentiary regime in product liability. The right to compensation under this instrument depends on the claimant's ability to prove the defectiveness of the product,[75] the damage suffered and the casual link between the two.[76] The defectiveness of the product "shall be presumed" in three cases, discussed further in this paper, but one of the three stands out: the case where there is evidence of the defendant's failure to comply with an obligation to disclose relevant evidence at their disposal.[77] The causal link between the defectiveness of the product and the damage "shall be presumed, where it has been established that the product is defective" and the harm caused is "of *a kind typically consistent* with the defect in question."[78] If due to technical or scientific complexity, the claimant experiences difficulties in proving defectiveness, the causal link or both, they can be presumed if the claimant

---

[72] *Id*., pt 30.
[73] *Id*., pt 31 (emphasis added).
[74] *Id*., pt 34 (emphasis added).
[75] *Id*., Art. 6 defines the notion of 'defectiveness' as failure to provide the safety which the public is entitled to expect, considering: the presentation of the product, including the instructions for installation, use and maintenance (a); the reasonably foreseeable use and misuse of the product (b); the effect of the product of any ability to continue to learn after deployment (c); the effect on the product of other products that can reasonably be expected to be used together with it (d); the moment in time when the product was placed on the market or put into service or, where the manufacturer retains control over the product, the moment when the product left the manufacturer's control (e); product safety requirements, including safety-relevant cybersecurity requirements (f); any intervention by a regulatory authority or by an economic operator (g), the specific expectations of the end-users for whom the product was intended.
[76] *Id*., Art. 9(1).
[77] *Id*., Art. 9(2). The duty to disclose evidence is enshrined in Article 8 R-PLD which states that national courts are empowered, upon request from the claimant "who has presented facts and evidence sufficient to support the plausibility of the claim for compensation, to order the defendant to disclose relevant evidence that is at its disposal" (Art. 9(1)). To determine if the disclosure is proportionate, national courts shall "consider the legitimate interests of all parties, including third parties concerned, in particular in relation to the protection of confidential information and trade secrets within the meaning of Article 2, point 1, of Directive 2016/943" (Art. 8(3)).
[78] *Id*., Art. 9(3), emphasis added.

gives "sufficiently relevant evidence" which shows that "the product contributed to the damage"[79] and "it is likely that the product was defective or that its defectiveness is a likely cause of the damage, or both."[80]

On the surface, these provisions do seem to lighten the burden for the claimants by conveniently setting out presumptions of defectiveness and/or causality. However, they are not - what we called in previous work[81] - *prima facie* presumptions *i.e.* facts held as established without prior evidence (like the presumption of innocence, for example). For the presumptions in the R-PLD to be established, the claimants carry the burden of establishing the basic facts (*indicia*) which if sufficient may, indeed, warrant the presuming of defectiveness and/or causality.

G. Structure and Outline of Main Arguments

To determine if and how explanations pertaining to AI output (as examined in connection to Explainable AI - XAI) can or should support explanations pertaining to the causal links between AI systems and harms suffered, we will follow, as *fil rouge* throughout this paper, the notion of *accuracy*. The inquiries that will frame our analysis are the following: does the accuracy of causal explanations in the field of AI liability *depend on* the accuracy of explanations pertaining to AI outputs? In the affirmative, which components should those explanations have, in order to be viewed as 'accurate'?

With accuracy in the backdrop, **Section 2** will lay down the analytical framework for the remainder of this paper, by focusing on the *type of knowledge* that explanations (*tout court*) provide and the standards that they respond to, in view of achieving accuracy or - at least - plausibility. Against the backdrop of various strands of history and philosophy of science, we will argue that, unlike 'scientific' knowledge (or 'knowledge proper'), explanations are held against *lower standards* of verifiability and accuracy, *believability* (in the eye of the explainee) being the criterion that truly sanctions - what scholars have called - the *goodness* of explanations (**Sub-Section 2.1.**). We will then go on to explore the 'goodness' conditions applied to explanations pertaining to causality in law (**Sub-Section 2.2.**). Since the purpose of *causal explanations* is to allow a competent authority (usually a court) to induce causation from series of correlations (*i.e*. positive associations between a conduct and a harm), the *evidence given*, as well as the criteria applied in its assessment are of utmost importance. Indeed, in cases where the cause-harm link is not self-evident or easily discernable, the type, probative value and relevance of the items of evidence given will play a major role in the mapping out of the stages that form the chain of causality which connects a wrongful and/or unlawful act to a damage.

With explanatory accuracy (*tout court* and in liability law) thus canvassed, **Section 3** will analyze how that concept relates to AI output. To do so, it will examine two sets of accuracy conditions: those applied to automated decisions/predictions and those applied to explanations of automated decisions/predictions. The first series of conditions will be our point of focus in **Sub-Section 3.1**. Bearing in mind the scholarship - explored in Section 2 - on the conditions for valid knowledge-construction, we will seek to determine if the 'knowledge' produced by non-human 'knowers'

---

[79] *Id*., Art. 9(4)(a).
[80] *Id*., Art. 9(4)(b).
[81] Ljupcho Grozdanovski, *La présomption en droit de l'Union européenne* (Anthémis, 2019).

presents any specificity (in terms of how it is formed and when it is 'accurate') in the context of traditional epistemology. Against this backdrop, we will raise the issue of the human knowability of AI output and critically address (and assess) the assumptions we make when we seek to explain automated decisions and predictions which are, partially or totally, inscrutable (and therefore, unknowable) by humans. Based on our exploration on the 'epistemic status' of AI output, **Sub-Section 3.2.** will examine the second series of the abovementioned conditions *i.e.* the accuracy standards for explanations pertaining to AI output. Going by the object of those explanations (*i.e*. the thing explained), we will focus our attention, first, on so-called *ad hoc* explanations (relative to the standards observed *a priori* in the inception of AI systems), second, on so-called *post-hoc* explanations (relative, in essence, to the reasoning patterns underlying harmful automated decisions and uncovered *ex post i.e*. once those decisions have been made). Our analysis of the 'accuracy' criteria applied for each of those two types of explanations will then allow us to critically examine the EU's regulation on AI liability and (finally) address the following issues: 1. whether said regulation - seeking to define systems of adjudication that would not leave litigants without effective judicial protection - allows for the adducing of evidence which can support *ad hoc* explainability, *post hoc* explainability or both; 2. whether the type of explanation required under said regulation takes into account what the *litigants themselves flag as necessary* for the purpose of making their views known and effectively participating in the adjudication of their disputes. Since the EU's AI liability regulation is not yet binding, there is no caselaw which can allow us to map out the procedural needs (in terms of evidence and explanations) that litigants have in disputes dealing with AI-related harm.

In **Section 4**, an examination of the evolving caselaw on AI liability, predominantly in North America, is presented to highlight pertinent procedural (and explanatory) needs. This analysis aims to delineate the types of understanding sought by litigants and courts in disputes related to AI. Drawing insights from specific, relevant cases, it becomes evident that the sought-after understanding in these disputes primarily revolves around two aspects. First (covered in **Sub-Section 4.1.**), there is a focus on the *accuracy of a given AI output*. The procedural concern here centers on whether there is sufficient evidence to ascertain the accuracy or inaccuracy of an automated decision. Second (explored in **Sub-Section 4.2.**), the attention shifts to the rationale justifying human reliance on the - potentially inaccurate and harmful - AI output. This raises the question of the *motives* having led a human agent to believe that an AI decision was accurate and, consequently, trustworthy. Our analysis of the emerging AI liability caselaw will allow us to identify *two trends* on the components of causal explanations: 1. XAI *is* integral to those explanations; 2. XAI should - ideally - be 'full' *i.e*. *ad hoc* and *post hoc*. Based on these conclusions, we will critically assess the EU's AI liability regulation which, from a procedural perspective, seems to restrict the scope of evidentiary debates in AI liability cases to *ad hoc* explainability only.

In our examination of the AILD and R-PLD, with a specific focus on the claimants' entitlement to seek evidence disclosure, the primary finding, highlighted in **Sub-Section 4.3.**, is that the evidence authorized under said instruments mainly supports *ad hoc* explanations. It reveals whether technical standards, particularly those outlined in the AI Act, were complied with in advance. Notably, there is an absence of provisions allowing litigants to receive *post-hoc* explanations, *i.e*. explanations on how a system *concretely* made a given harmful decision.

This will allow us to, in **Section 5**, express our criticism of the AILD and R-PLD. Based on our analysis of the emerging AI liability caselaw, our critique will translate to a plea to interpret (or amend) these instruments so that they may include the *procedural abilities* (to give and receive evidence and explanations) that the litigants in said caselaw flagged as necessary. For victims of AI-related harm, we will argue that *ad hoc* explainability is not enough: what claimants aim to understand when seeking compensation are the components of the causal link between an AI system's functionalities, reasoning patterns and output, on the one hand and a harm suffered, on the other hand. For that purpose, *post-hoc* explainability is paramount. Alternatively, for defendants, we will inquire if the AILD and R-PLD allow them to effectively exercise their right to defense. This inquiry is motivated by the fact that both the AILD and R-PLD charge the defendants with providing evidence (to claimants), but neither raises the question of whether the defendants themselves might need evidence to be disclosed (say, expertise) so that they may defend themselves more successfully against the claimants' allegations.

Drawing from the aforementioned points, our final remarks regarding the correlation between evidence, explanation, and procedural fairness are presented in **Section 6**.

I. ACCURACY OF EXPLANATIONS *TOUT COURT*

In adjudicatory contexts, evidentiary explanations are meant to provide *understanding* of the disputed facts, which explainees (such as courts and juries) are likely to view as accurate or at least, convincing. The epistemic question here is, of course, that of the *criteria* that ought to be met for explanations to qualify as 'accurate.' The ambition of this Section is to uncover those criteria, determine how they translate into law and lay the conceptual framework within which the remainder of this paper can take shape. To do so, **sub-Section 2.1.** will go back to the 'source' and delve into the concept of accuracy in connection to *scientific knowledge*, as the epistemic template (*genus*) for the concept of *explanatory* knowledge. Against the backdrop of our analysis of the knowledge/explanation kinship, **sub-Section 2.2.** will zoom in on *causal explanations* sought for the purpose of adequately representing and proving causality in law.

A. Scientific Knowledge, A Model for Explanatory Knowledge

Aspiring to be agnostic, epistemology abhors bias, one of its longstanding battles having been to 'cleanse' knowledge from preconceptions, beliefs and representations residual in the knowers' minds.[82] Knowledge - Bonderup Dohn insists - should "not just be employed as a black box term or be characterized only in its correlation with, for example, psychological states or social relations without being given an explicit analysis as regards its nature."[83]

But what is 'objectivity'? Minazzi posits that the term 'objective' "refers, in the first instance, to what exists as an object or to what possesses an object or, again, to

[82] Cartesian doubt is echoed here which consists in insisting that "the existence of a thought does not in itself guarantee the existence of what it purports to be a thought of." See Peter Winch, *Spinoza on Ethics and Understanding*, *cit. supra*, at 4.

[83] Nina Bonderup Dohn, "Epistemology in Investigating Knowledge: 'Philosophizing with" (2011) 4 *Metaphilosophy*, 431, at 431.

what belongs to an object."[84] In a similar vein, 'objective' - he argues - indicates both "everything which appears to be valid for everyone, and what appears to be independent of the subject, as well as everything which is 'external' with respect to consciousness of thought and, last but not least, everything which is found to comply with certain rules or methods."[85]

Minazzi's observations allow us to assert that, in the heart of any knowledge-construction endeavor lies the question of whether objectivity translates to *a*-subjectivity *i.e.* subject-independence (the 'subject' here being the person who acquires knowledge, not the person to whom knowledge is communicated).

A brief historic overview of epistemology reveals an original *penchant* for objectivism, characterized by the search for methods meant to 'cut off' (as it were) the knowledge of the outside world from the knower's inner world (**Sub-Section 2.1.1.**). This current was, however, contrasted by subjectivism (**Sub-Section 2.1.2.**) which was eventually - and, in some cases, reluctantly - accepted as unavoidable. Scholars came to realize that, try as they might, 'valid' knowledge could never be fully divorced from belief; an observation which applies *a fortiori* to explanations for one simple reason: their accuracy does not solely depend on the explainor's epistemic and communicative competence. It also, if not mostly, depends on the explainee's *ability to understand* the gist of the explanation given.

1. The Ideal(ized) Objectivism

Enlightenment philosophical traditions - namely Newtonian physics and Kantian transcendental philosophy[86] - gave us the analytical benchmarks we turn to, in order to develop our normative understanding of what 'pure' science is. The methods of scientific discovery and the criteria used for the validation (or invalidation) of knowledge began in the natural sciences, subsequently shaping the epistemology in the social sciences,[87] including law (especially, the law of evidence).[88]

Throughout its evolution and the constant fine-tuning of the criteria for 'true' or 'valid' knowledge, epistemology maintained its original posture of agnosticism, the idea being that knowledge ought to include *belief-independent accounts* of the world and not be 'corrupted' by the knower's *representations* thereof (**A**).

This *penchant* for objectivism is particularly visible in the verificationist strands on explanations. However superficial they might seem - compared to the cognitive depths that knowledge proper aspires to reach - explanations remain *fact-correspondent* that is, pertain to an *object of explanation* that is material, tangible and verifiable (**B**).

---

[84] Fabio Minazzi, *Historical Epistemology and European Philosophy of Science* (Springer, 2022), at 3.
[85] *Id*., at 3-4.
[86] See e.g. Michael Friedman, "Newton and Kant on Absolute Space: From Theology to Transcendental Philosophy" in Michel Bitbol, Pierre Kerszberg, Jean Petitot (ed.), *Constituting Objectivity. Transcendental Perspectives on Modern Physics* (Springer, 2009), 35-50.
[87] For the translation of the 'scientific method' in sociology (the seminal figure of which is, of course, Durkheim), see Enzo Di Nuoscio, "L'individualisme méthodologique comme méthode scientifique: théorie de la rationalité, explication causale, herméneutique" (2020) 70-1, *L'année sociologique*, 129.
[88] The process of giving and assessing legal evidence has been labelled as 'courtroom epistemology.' See Baosheng Zhang, Jia Cao, David R.A. Caruso, "The Mirror of Evidence and the Plausibility of Judicial Proof" (2017) 21 *Int'l J. of Evidence & Proof*, 119, at 123.

a. The Belief-Independence of Knowing

'Proper' knowledge - often synonymized with 'scientific' knowledge[89] - is meant to somehow capture the essence of the portions of reality it pertains to. In its purest, most idealized flavour, it is meant to uncover the "exceptionless laws"[90] that govern the phenomena that fall in the scope of our experiences of reality. In the backdrop of this ideal, it is not surprising that objectivity has been historically fetishized, fostering hostility toward the belief-ridden persona of the knower, 'belief' being usually seen as an irrational creed, held "for a reason which is preposterous or for no reason at all."[91]

Is it possible for someone to pursue knowledge of reality without being emotionally and cognitively tainted by their beliefs? Isn't knowledge itself a set of - as Keynes put it - *rational* beliefs[92]? The belief/knowledge interplay has been a recurring theme in savant circles, which offered varying views on the *posture(s) of the knower* and the *models of reality*. Putnam seminally argued that three important (meta) traditions addressed these issues: "the extreme *Platonist position* which posits non-natural mental powers of directly 'grasping' forms (…) the *verificationist position* which replaces the classical notion of truth with the notion of verification or proof and there is the *moderate realist position* which seeks to preserve the centrality of classical notions of truth and reference without postulating non-natural mental powers."[93] Each meta tradition has given way to numerous sub-strands, the detailed accounts of which fall - alas - outside of the scope of this paper. For the sake of brevity, let us refer to Minazzi's work on *epistemic objectivity*,[94] a point on which he sought guidance in Kant's work.

Kant's brilliant philosophy arguably made two major contributions to the ways in which we understand and construct (objective) knowledge. First, that discovery of

---

[89] Considering traditional epistemology is characterized by three central notions namely knowledge, belief and doubt, securing a level of stability of knowledge appeared as a process of responding to skepticism while, at the same time, creating models of 'valid' epistemic models (*i.e.* models able to reliably deliver knowledge). See Vincent F. Hendricks, John Symons, "Where's the bridge? Epistemology and Epistemic Logic" (2006) 128 *Philosophical Studies*, 137, at 138-139.

[90] We borrow here Putnam's expression used in her comment of Quine's (post-Kantian) view on - what she called - *analyticity*, essentially derived from Kant's concept of analytic judgments. See Hilary Putnam, *Realism and Reason. Philosophical Papers* (vol. 1, CUP, 2010), at 89.

[91] John Maynard Keynes, *A Treaties on Probability* (Macmillan & Co., 1921), at 10.

[92] "Knowledge of a proposition - Keynes writes - always corresponds to certainty of rational belief in it and at the same time to actual truth of the proposition itself. We cannot know a proposition unless it is in fact true." John Maynard Keynes, *A Treaties on Probability*, *cit. supra*, at 11. Keynes' observation is interesting for mainly two reasons. On the one hand, he views knowledge as propositional (knowledge consists of propositions about reality). On the other hand, he dissociates certainty and truth as if to distinguish a justified belief of accuracy (certainty) from accuracy *tout court* (truth). We will not further discuss this distinction, however interesting and relevant it may be for our discussion on the epistemic ideal of objectivism and the epistemic 'tolerance' of subjectivism. Keynes wrote a seminal (though a bit dated) work on probability and defined knowledge-as-certainty so that he could then delve into the concept of probability. However brilliant, he is not in the forefront of strands on explanation, which are the main focus of this paper.

[93] Hilary Putnam, *Realism and Reason. Philosophical Papers*, *cit. supra*, at 1 (emphasis added).

[94] Minazzi raises the longstanding and complex question of whether scientific knowledge can really be value-free? The assumption here is that science is apolitical and acultural knowledge production activity. However, Minazzi ultimately concludes that scientific knowledge and its production are rooted in "a stratified social reality that may produce different images of the human knowledge itself. See Fabio Minazzi, *Historical Epistemology and European Philosophy of Science*, *cit. supra*, at 121.

knowledge is usually not serendipitous, but the result of highly protocolized epistemic processes.[95] Second, under Kant's influence, we place the inferences drawn from our discoveries on a "new heuristic plane of *transcendentality*, by which Kant constructs the overall theoretical framework of his epistemological meta-critic reflection, deeply innovating not only the whole concept of knowledge, but also the style and modes of human rationality."[96] Scientific knowledge is 'scientific' because science is "always capable of thinking its object by constructing it through a plastic critical interplay of continuous comparison with the experimental dimension." [97] In other words, knowledge is produced *within* the confines of already existing conceptual frameworks, where well-established (and constantly perfected) sets of epistemic competences are deployed.[98]

In addition to Minazzi, other contemporaries expressed similar intuitions. Latour seminally expressed the view of trials and experimentation as being "ritual frameworks" with value hierarchies that 'actants' (which include humans as well as, say, microbes) obey in the fabrication of 'scientific facts.'[99] In this context, knowledge proper can be understood as an *adept belief*[100] - the word 'belief' again! - for which an epistemic community considers there to be *sufficient reasons* to hold it as true, at least until more conclusive, belief-dispelling evidence is brought forward.[101]

Though Kant - and others - molded our modern understanding(s) of scientific epistemology, pushing us to sharpen our intuition on what *true* science is, the subjectivism/objectivism dilemma was not altogether effaced from epistemic discourse. To this day, an opposition remains between *materialists* who view facts as the sole gateways to agnostic truth and *mentalists* for whom, knowledge formation carries the

---

[95] Duede writes: "scientists do not *design* the physical processes. Rather, they, as it were, *discover* them. With theory mediated instruments, nothing is out of our hands." Eamon Duede, "Instruments, agents, and artificial intelligence: novel epistemic categories of reliability" (2022) 200 *Synthèse*, 491, at 501.

[96] Fabio Minazzi, *Historical Epistemology and European Philosophy of Science*, *cit. supra* at 11 (emphasis added).

[97] *Id*., at 12.

[98] See e.g. Susanne Mantel, ''Acting for reasons, apt action and knowledge'' (2013) 190 *Synthèse*, 3865, at 3873.

[99] See Kyle McGee, *Bruno Latour: The Normativity of Networks* (Routledge, 2014), at 4.

[100] John Turri, ''Manifest Failure: The Gettier Problem Solved'' (2011) 8 Philosophers, 11 *cit. in* John Greco, "A (different) virtue epistemology" (2012) 1 *Phil'y & Phenomenological Res.*, 1, at 9.

[101] This type of belief-forming epistemic practices (and the virtues or values associated with those) were examined by Sosa in his study on reflective knowledge (essentially focused on the reliability and criteria used to label something as 'knowledge') as opposed to 'animal' knowledge,' which is mostly perceptual, experiential with no ambition to systematize a set of protocols and procedures meant yield Sosian 'apt' beliefs. See Ernest Sosa, *Reflective Knowledge: Apt Belief and Reflective Knowledge* (vol. II, OUP, 2009), at 135 seq.

imprint of the knower's 'mental states' (social contexts, background knowledge and preexisting values and beliefs).[102]

Epistemology's aspiration to cut the umbilical cord between knowledge proper and psychology has transpired into modern evidence scholarship.[103] Of course, the knowledge derived from legal evidence has never been held to the validity standards of science. Nevertheless, the requirement for objectivity has woven through schools of thought on evidence, which can be perceived as unrealistic. When litigants give evidence in a trial, they do so as adversaries, confronting their *versions of the disputed facts* with the goal of winning the case. The answer to the question 'what happened in a dispute?' is, in a way, doomed to be subjective since "it's not about truth, it's about who tells a better story."[104]

However, we mentioned earlier that legal evidence is a peculiar beast,[105] namely because the adducing of evidence should both *be fair* and serve *the purpose of fairness* (*i.e.* a fair resolution of a dispute). It is precisely because of this 'fairness constraint' that the epistemology of legal evidence has - heavily! - drawn inspiration from scientific discovery methods. The idea is that 'adequate' (*i.e.* impartial, politically, culturally and socially neutral and therefore fair[106]) administration of justice requires *some level* of objectivity in the ways in which facts are given and assessed. In this context, a law of evidence is typically meant to (at least minimally) define basic epistemic conditions under which litigants can debate facts and do so before an unbiased authority.

By defining the features of various types of evidence (admissibility, probative value, standards of proof) and the requirements for fact-appraisal (impartiality, legal expertise, fairness), a law of evidence does not establish a scientific discovery-type proceduralism conducive to measurable, verifiable and reproducible results. It does, however, provide a set of *procedural guarantees* meant to preclude evidential truths from depending on the whims of litigators, judges and juries. Those guarantees (mainly linked to the parties' equal opportunity to plead and the courts' independence) warrant,

---

[102] Scientific truths are essentially beliefs held as true or beliefs for which there are good, or valid reasons to accept as true. Beliefs stand so long as they are justified which, of course, begs the question of the conditions that warrant justifiability. In an outline of the main schools of thought within epistemology, Bishop and Trout distinguished three: foundationalism, coherintism and reliabilism. The first two are internalist theories of justification, in the sense that the 'justifiers' for holding a belief as true are accessible to the believer. Foundationalists - Bishop and Trout argue - hold that "many beliefs are justified in terms of their relations to other beliefs." This presupposes a set of basic beliefs that act as 'normative justifiers' of sorts and in reference to which subsequent beliefs are assessed. Coherentists are a spin-off from foundationalists: they also consider that what beliefs can be justified in terms of their relations to other beliefs, but coherentists deny the existence of basic beliefs. For reliabilists, the justifier is external: a belief is justified in case it is produced by a reliable belief-forming mechanism. See Michael A. Bishop, J. D. Trout, "Epistemology's search for significance" (2003) 15 *Journal of Experiential & Theoretical Artificial Intelligence*, 203, at 205.

[103] Modern evidence theory roughly includes the past 200 years of scholarship. See namely Douglas Walton, *Legal Argumentation and Evidence* (Penn. State. Univ. Press, 2002), at 106.

[104] Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication" (2018) 26 *AI & Law*, 345, at 347.

[105] See *supra*,

[106] As May put it, "procedural justice conveys the idea that everyone will be subject to and protected by the same rules. Each person is to be seen as equal before the law." Larry May, *Global Justice & Due Process* (CUP, 2011), at 13.

if not the certainty, at least the *expectation* that the law, impersonal and fair, will *deliver justice*.[107]

Our brief *exposé* on subjectivism and objectivism as debated among epistemologists and as taken over by procedural lawyers, sets the tone for our analysis of *evidentiary explanations*, the accuracy of which is also characterized by a quest for balance between independent (impersonal) standards and subjective beliefs. Against this backdrop, we can - finally - raise the questions we wish to address in this sub-section: 1. what is explanation?; 2. what is an accurate (good) explanation? Though straightforward answers are hardly possible, we will - in a pedagogical *élan* - distinguish two definitions, one we will call static (the act of explaining) the other, dynamic (the process of explaining).

b. The Fact-Correspondence of Explaining

In a static sense, an explanation is, in essence, an *interpretation of experience*: to explain is to provide meaning of specific objects[108] in understandable terms.[109] Of course, for explanatory interpretation to be possible, the object of explanation should be interpretable, interpret-*ability* being the feature of the object explained to acquire concrete meaning.[110]

In AI jargon, interpretability and explainability are often used interchangeably though Hauque *et al.* view them as conceptually distinct. Explainability, they argue, "means explaining the decisions made by machine models in a human-understandable form."[111] Alternatively, interpretability "is the explanation of how or why a model resulted in a particular prediction."[112] However plausible this distinction may be, we will consider, in the remainder of this paper, that *any* explanation (in the field of AI or not) is inherently interpretative. Indeed, to interpret an AI system's decisional process (what Hauque *et al.* call 'interpretability') is to provide the basis for an explanation of its output (explainability *stricto sensu*). Though there might be a semantic or a theoretical interest in distinguishing the two, for the purpose of this study, we will consider interpretability (*i.e*. a system's aptitude to be interpreted) as an epistemic

---

[107] There has been much debate on whether a law of evidence (as a consolidated *corpus* of rules framing evidentiary epistemology) can legitimately exist only if it is codified or it can also emerge from court practice. In an reductionist attempt, Wróblewski argued that a law evidence can be viewed as existing if it includes rules and principles which answer *four essential questions*: how does law distinguish between facts that require evidence from those that do not?; which evidence is admissible?; how is evidence assessed?; what is the role of evidence in the performance of (judicial) review?. See Jerzy Wróblewski, « La preuve juridique : axiologie, logique et argumentation » *in* Chaïm Perelman, Paul Foriers (ed.), *La preuve en droit : études* (Bruylant, 1981), 331, at 338.

[108] William Franz Lamberti, "An overview of explainable and interpretable AI' in Feras Baratesh, Laura Freeman (eds), *AI Assurance. Towards Trustworthy, Explainable, Safe and Ethical AI* (Elsevier, 2022), 55-123, at 57.

[109] Ricardo Guidotti, Anna Monreale, Dino Pedreschi, Fosca Giannotti, "Principles of Explainable Artificial Intelligence" *in* Moamar Sayed-Mouchaweh (ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications* (Springer, 2021), 9, at 12.

[110] *Ibid*.

[111] Bahalul Haque, Najmul Islam, Patrick Mikalef, « Explainable Artificial Intelligence (XAI) from a user perspective : A synthesis of prior literature and problematizing avenues for future research" (2020) 186 *Technological Forecasting & Social Change*, 1, at 2-3.

[112] *Ibid*.

precondition for explainability proper (*i.e*. interpretation given on the system's functionalities and decisional/predictive processes).

As with any type of interpretation, the risk with explanations is that of *mis*interpretation. Badea and Artus[113] call this the interpretation problem (IP). The threat of IP calls for caution because virtually any real-world occurrence can be interpreted in infinite and unspecifiable ways. In the field of AI, the IP arises - Badea and Artus argue - because of the possibility that "a highly advanced machine may find novel interpretations of the rules that we give it, interpretations which are not incorrect, in that they can be seen as valid interpretations of the rule, but which are inappropriate in that we do not approve of them."[114] As a mitigation strategy to the IP (in the field of AI, and in general), explanation theorists sought to define *basic accuracy criteria* which can be clustered in two families: those - leaning toward objectivism - that support explanatory *fact-correspondence* or *facticity* and those - subjectivist-prone - that support *understandability*.

Regarding facticity, the theoretical referent we will use is the so-called *correspondence* theory of truth, which upholds the view of "agreement or correspondence between a statement and the so-called facts or reality."[115] It should be mentioned that correspondence theory does not eradicate subjectivism altogether; its gist consists in preferentially using perceptive reality as 'the' referent for the validity of propositions made about that reality. For instance, if one wishes to know if they may justifiably assert that snow is white, they ought to see the color of snow falling (*i.e*. turn to reality to verify the truth or falsity of the 'snow is white' statement).

That explanations should be in accord with tangible facts does not raise any particular controversy: if this was not the case, there would be next to no difference between explaining and the "narrative techniques of imaginative writers."[116] In adjudicatory contexts, explanations' *rattachement* to reality is paramount precisely because it *enables verification*: when courts are called to resolve disputes, they strive to acquire, from the parties, *accurate* knowledge of facts so that they may draw relevant conclusions on important legal (and by that, social and political) issues like guilt or liability.[117]

---

[113] Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents," *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (2022), SGAI-AI, AI XXXIX, 124.

[114] *Id*., at 125.

[115] Carl G. Hempel, "On the Logical Positivists' Theory of Truth," *in* Richard Jeffrey (ed.), *Selected Philosophical Essays* (CUP, 2012) 9, at 9. Hempel opposes the correspondence truth theory to the coherence theory of truth, according to which "truth is a possible property of a whole system of statements." See *ibid*. Exploring the relevant ways in which correspondence and coherentism are similar, complementary or opposed is beyond the scope of this paper. We refer to the correspondence-theory as a theoretical referent allowing us to make the following point: if explanations are taken as statements *about facts*, they ought to relate to those facts, 'facts' being understood as tangible, perceptible, verifiable *objects of experience*. The choice of correspondence theory is important because it accepts that facts are facts, regardless of whether there are propositions made about them (this is the coherentist thesis according which - if we were to vulgarize it - there are no facts *per se*, only propositions about facts). Whether the explanation-fact correspondence is well-established (adequate or credible) is an issue of assessing the *conditions* under which truth-as-correspondence can stand as acceptable (and therefore accurate). These conditions will be discussed further in this paper.

[116] Simon Stern, "Factuality, Evidence and Truth in Factual Narratives in the Law," *cit. supra*, at 391.

[117] *Id*., at 392.

Referring to Di Bello's probabilistic analysis of criminal trials,[118] Urbaniak stressed that "the relationship between evidence and (evidentiary) narratives goes both ways: from the evidence to the narratives and from the narratives to the evidence."[119] There is something intuitively convincing about this interplay: evidence is both the *foundation* for a narrative about facts and the *standard* against which the validity of the truth of that narrative is assessed. While narrations - Urbaniak writes - play a "crucial role in the account, their relation to evidence and their factual support is also in the focus, hopefully susceptible to a more precise probabilistic analysis."[120] In laymen's terms, when we say that something is true or false in an adjudicatory context, 'something' is usually a state of fact.

The need for explanations to be fact-correspondent implies that they are *context specific*[121] and *factive*.[122] The role of context in assessing the explanatory goodness (understood here as a 'thin' concept of accuracy) will be discussed further. At this stage, may it suffice stressing that facticity is, indeed, the unavoidable but not exclusive referent for the assessment of said goodness. We do not *explain* gravity simply by advising someone to drop a pen. By dropping the pen, they *experience gravity*, but do not gain understanding on what it is and why it works the way it does. All explanatory contexts include an actor who ultimately says 'yay or nay' on the accuracy/plausibility reached (or not) by the explanation given. Enter the figure of the explainee.

2. The Unavoidable Subjectivism

As mentioned earlier, epistemology's aversion to belief has been somewhat 'diluted' in contemporary scholarship. Rarely does a fact speak for itself, declaring - as it were - a truth about the world irrespective of an observing knower's perceptions and beliefs. For example, regardless of how one feels about water's boiling point, it will always be reached at 100 °C. Even propositions (like, 'the sky is blue') which we as laymen view as uncontroversial, have triggered erudite debates on the conditions under which those propositions could be held as true (obviously, because the sky is not always blue)… Our point is the following: any type of knowledge is to some degree belief-dependent: a proposition (hypothesis, theory, explanation…) about a state of facts is true to the extent, and so long as relevant expert and/or non-expert communities believe it to be. In explanatory contexts, *believability* does, indeed, appear to be the apex standard for explanatory accuracy (**A**), assessed by the explainee in reference to the context in which they receive a specific explanation (**B**).

a. Believability as Proxy for Explanatory Accuracy

Delivering understanding, as the purpose of any explanation, allows us to tackle the above-mentioned *dynamic definition* (*i.e*. explaining as a process). Explanations are

---

[118] Marcello Di Bello, *Statistics and probability in criminal trials*, Ph.D. Thesis (University of Stanford., 2013).
[119] Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication" (2018) 26 *Artif. Intell. Law*, 345, at 348.
[120] *Ibid*.
[121] Michael Ridley, "Explainable Artificial Intelligence (XAI)" (2022) 2, *Information technology and libraries*, 1, at 3.
[122] Andrés Paez," The Pragrmatic Turn in Explainable Artificial Intelligence" (2019) 3 *Minds and machines*, 441, at 454.

communicative acts: in most cases, something is explained by someone, to someone.[123] Bearing in mind standard knowledge-construction theories, we tend to place our focus on what the explainer ought to do to deliver clear and accurate information. But communication is a two-way street, the level of comprehensibility of the explanation given depending also (if not predominantly) on the explainee's *ability to understand*.

This ability is largely shaped by the explainee's prior knowledge and experience. For instance, a flat-earther will likely discard the many photos taken from space showing the Earth's spherical shape. Those photos would presumably be dismissed as untrustworthy evidence in the face of a person's unwavering belief that the Earth is, in fact, flat. The point we seek to make through our flat Earth example is the following: though anchored in facts, explanations will always be viewed through the lens of their addressees' beliefs and because of this, they will likely fall on biased ears.[124] In the context of AI, the trustworthiness of explanations (pertaining to, say, the probability that a system develops a gender bias) will largely depend on whether explainees look favorably on AI to begin with.[125]

In examining the explanatory process through the vantage point of communication, Ridley[126] highlighted three features all explanations share. First, they are *contrastive*: when people want to know the 'why' of something, they "do not ask why event *p* happened, but rather why event *p* happened *instead* of some event q."[127] Second, they are *selected*: people are adept at "selecting one or two causes from a sometimes infinite number of causes to be *the* explanation."[128] Third, explanations are *social*: "they are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs."[129]

The fact that an explanation is given *to* someone places, on the explainer, the duty to deliver, to the best of their ability, adequate understanding of the object explained, 'adequate' explanations being, in essence, those that manage to warrant *believability*. It can even be argued that believability is for explanations what accuracy *strico sensu* is for knowledge proper. According to Paez, this believability-as-proxy-

---

[123] Denis J. Hilton, "Conversational processes and causal explanation" (1990) 1, *Psychological Bulletin*, 65, at 65.

[124] This is of course not the least bit surprising, considering that comprehensibility is, typically, a matter of making associations between what a person views as true and what is, to them, new information. As Moehring *et al.* stress, the 'comprehension construct' is a process of developing mental representations, by which prior long-term knowledge is incorporated with the available information." See Anne Moehring, Ulrich Schroeders, Benedikt Leichtmann, Oliver Wilhelm, "Ecological momentary assessment of digital literacy: Influence of fluid and crystallized intelligence, domain-specific knowledge, and computer usage," (2016), 59 *Intelligence*, 170, at 171.

[125] In Vered *et al.*'s excellent work, we find interesting empirical studies (and corresponding inferences) on the interrelationship between explainability of AI and reliance on automated decisions. Based on several empirical studies in radiology, the authors conclude that local and global explanations tend to decrease over-reliance, decreasing the explainees' automation bias. See Mor Vered, Piers Douglas Lionel Howe, Tim Miller, Liz Sonenberg, "The effects of explanations on automation bias" (2023) 322 *AI*, 103952.

[126] Michael Ridley, "Explainable Artificial Intelligence (XAI)," (2022) 41-2, *Inf. Tech'y & libraries*, 1, at 4.

[127] *Ibid*.

[128] *Ibid*.

[129] *Ibid*.

for-accuracy is due to the fact that understanding is not strictly speaking knowing[130] which *a fortiori* suggests that explaining is not strictly speaking discovering.

Paez's intuition is on point. As mentioned earlier, (scientific) knowledge is 'knowledge' because it is "supported by protocol statements"[131] and accepted as such by communities who share the same epistemic competences. Echoing verificationist[132] and Latourian views on knowledge-construction, scientific experiences are often highly proceduralized, the threshold of accuracy (that is, justified acceptance of beliefs) being usually quite high.[133] Because of this, scientists can test accepted beliefs on a continued basis, constantly revisiting the reasons why a theory should remain acceptable.[134] The 'knowledge' explanations provide is of a slightly different kind. They are not issued from discovery *per se*. They rather provide "a kind of packaged summary of the relevant events; and if successful, this summary allows us to make appropriate inferences of the situation."[135] For an explanation to be qualified as 'good,' the golden rule seems to be '*know thy audience*.'[136]

But here, an interesting question emerges: how do facts support explanatory believability? If believability on the side of the explainees is, indeed, the workable variant of accuracy applied in explanatory contexts, do we still need evidence to support the explanation's fact-correspondence? In other words, are explanations 'accurate' only when the explainees believe them to be so, regardless of the interpretations warranted by facts? Subjectivism again rears its ugly head and its 'threat'[137] should not be underrated, given that - as mentioned earlier - explanations, like any form of knowledge, are not pulled out of thin air, but must have *some* anchoring in reality. We thus circle back to the debate previously canvassed on subjectivism and objectivism as the two points of oscillation of modern conceptions of epistemic accuracy. Explanations have not been spared from this debate, as confirmed by representatives of several '-ism' strands.

---

[130] Andrés Paez, "The Pragmatic Turn in Explainable Artificial Intelligence" (2019) 29-3, *Minds & machines*, 441, at 453.
[131] Carl G. Hempel, "On the Logical Positivists' Theory of Truth," *in* Richard Jeffrey (ed), *Selected Philosophical Essays* (CUP, 2012), 9, at 9.
[132] Hilary Putnam, *Realism and Reason. Philosophical Papers*, *cit. supra*, at 89.
[133] A nuance and a clarification should be provided here. The nuance is that no knowledge, be it scientific of explanatory, is absolutely accurate. As Hempel rightly put it, "nowhere in science will one find a criterion of absolute unquestionable truth." See Carl G. Hempel, "On the Logical Positivists' Theory of Truth," *cit. supra*, at 16. The just like the process of scientific discovery is protocolized, the accuracy is as well, in the sense that, what is to be viewed as valid (in the sense of accurate) knowledge is a matter of convention among the members of epistemic communities. This is the effect of (conventionally agreed upon) epistemic norms, which "identify the conditions under which someone should or should not believe, do, or feel something." See Clayton Littlejohn, "Objectivism and Subjectivism," *in* Veli Mitova (ed.), *The Factive Turn in Epistemology* (Cambridge University Press: 2018), 142, at 142.
[134] This is what Minazzi calls "radical critical discussion." See Fabio Minazzi, *Historical Epistemology and European Philosophy of Science*, *cit. supra* at 154.
[135] Laura Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference from Explanation" (2022) 151-7, *J. of exp'l psy'y*, 1481, at 1482.
[136] Michael Ridley, "Explainable Artificial Intelligence (XAI), Information technology and libraries," *cit. supra*, at 3.
[137] Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication," *cit. supra*, at 347.

Evidential *explanationism* is associated with Allen and Pardo[138] who opposed the prevailing probabilistic current in contemporary evidence scholarship,[139] Bayesian probabilities[140] being a prominent school of thought within that scholarship. For probabilists, accuracy is usually function of: 1. the quality of the items of evidence presented in support of a given claim and 2. the individual (and numerically represented) probative value given to each item (e.g. A is true with probability of 0.52/1).[141] In contrast, explanationists view evidence as discursive and "inherently comparative - whether an explanation satisfies the standard depends on the strength of the possible explanations supporting each side."[142] Because of this, explanationist accuracy is, indeed, synonymous with believability: it is not about meticulously measuring probabilities, but about the (non-quantifiable) levels of persuasiveness an explanation can generate. This is understandable, considering how impractical it is to expect factfinders to "actually attach probabilistic numbers to each probability at issue in litigation."[143]

Evidentialism also leans toward a more subjectivist view of explanatory accuracy. Seminally represented by Conee and Feldman,[144] the gist of this current is that "the epistemic justification of a belief is determined by the quality of the believer's evidence for the belief."[145] In truth, Connee and Feldman (re)state longstanding strands in epistemology on the conditions under which beliefs can be justifiably held as valid (*i.e*. taken as true until rebutted). Thinkers like Locke, Hume, Reid and Bentham have long ago "championed or at least anticipated evidentialism."[146] Much like explanationists, evidentialists do not suggest some metric system that would allow us to numerically represent the truth value of explanations. They remain 'subjectivists': Conee and Feldman even called themselves mentalists,[147] positing that the justification

---

[138] Ronald J. Allen, Michael S. Pardo, "Relative plausibility and its critics" (2019) 23-1/2, *The Int'l J. of Evidence & Proof*, 5. The authors argue that the explanationist approach to legal evidence presents advantages that the probabilistic approach does not offer, namely "(1) the need to assign number values to compare the standard of proof; (2) lack of fit between the probabilistic theory and how fact-finders actually evaluate and reason with evidence; (3) inconsistency with legal doctrine and jury instructions (the conjunction problem); and (4) inconsistency with regard to the policy goals underlying standards of proof." See *id*., at 17.

[139] Under this probabilistic view, evidence is represented as *measurable* assessment of the likelihood of the disputed facts. See, for instance, Paul Horwich, *Probability and Evidence* (CUP, 2016), at 100 seq.

[140] Urbaniak provides a concise summary of Bayesian theory. Standard Bayesian epistemology "represents degrees of belief (also known as credences) by real numbers. Degrees of belief of an ideally rational agent, on the standard view, should satisfy the standard axioms of probability: probability should take values between 0 and 1 inclusive, logically impossible events get probability 0, logically certain events have probability 1, and the probability of the union of finitely many disjoint events is the sum of their individual probabilities (in the context of this paper, whether this holds also for infinite unions will not come up)." Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication," *cit. supra*, at 353. For an analysis of the application of Bayesian theory in the field of legal evidence, see, *inter alia*, Terence Anderson, David Schum, William Twining, *Analysis of Evidence* (CUP, 2009) at 246 seq.

[141] Johan B. Gelbach, "It's all relative: Explanationsim and probabilistic evidence theory" (2019) 1-2 *The Int'l J. of Evidence & Proof*, 168, at 171.

[142] Ronald J. Allen, Michael S. Pardo, "Relative plausibility and its critics," *cit. supra*, at 15.

[143] *Ibid*.

[144] Earl Connee, Richard Feldman, *Evidentialism: Essays in Epistemology* (Oxford University Press: 2004).

[145] *Id*., at 83.

[146] Philipp Berghofer, *The Justificatory Force of Experiences* (Springer : 2022), at 69.

[147] Earl Connee, Richard Feldman, *Evidentialism: Essays in Epistemology*, *cit. supra*, at 99.

of a belief that evidence is true largely depends on the "totality of one's mental states"[148] which are "drawn from *our experiences* as points of interaction with the world."[149] Conscious awareness - they write - is how "we gain whatever evidence we have."[150] Consequently, "much of what we know about the causal structure of the world we infer from directly observing and interacting with it."[151]

In sum, 'accurate' explanations are believable based on a *level of coherence* between the evidence given by the explainor and the explainees' residual beliefs. What reinforces this coherence is the *context* within which explanations are given. Indeed, as in most real-life situations, for explanations too, context is everything.

b. The Benchmark for Believability: Context is Everything

Is there an independent standard against which explanatory believability can be assessed? A definitive answer is next to impossible to give. Generally, evidentialists allude to *shared or common experience* or - to be more exact - *conventional interpretations of reality*.[152] Scholars have called this the *justificatory role of experience*.

Common experiences form - to paraphrase Aristotle - the realm of *doxastic knowledge*:[153] not knowledge *per se*, but a form of 'common wisdom' derived from experiences shared within given communities. Doxa gives people a sense of normalcy, a state of affairs where certain facts (e.g. children born in wedlock are fathered by their mothers' spouses) are accepted as true because they are perceived as 'normal.' In the case of AI, *no one* would ask for an explanation on how an AI system became gender-biased, if that bias was not viewed as a deviation from what the explainees view as a normal state of reality. Such a bias would be perceived as an error, the conventional belief - though often dispelled - being that unfair biases have no place in a world where equality should be the social and legal norm.

The concept of normality is a can of worms in its own right, usually defined through two main versants: *descriptive* (normality derived from the repetition of events) and *prescriptive* (state or conduct resulting from convention).[154] In causal contexts, Kirfel *et al*.[155] confirm through empirical data what Hart and Honoré[156] had previously claimed in legal theory - people tend to designate *abnormal events* as causes of harm:

---

148 Philipp Berghofer, *The Justificatory Force of Experiences*, *cit. supra*, at 70.

149 Earl Connee, Richard Feldman, *Evidentialism: Essays in Epistemology*, *cit. supra*, at 87.

150 *Ibid*.

151 Lara Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference From Explanation" (2022) 7 *Journal of Experimental Psychology*, 1481, at 1482.

152 We allude here to Michalski's definition of experience as the totality of information generated in the course of performing some actions. See Ryszard S. Michalski, "Inferential Theory of Learning as a Conceptual Basis for Multistrategy Learning" (1993) 11 *ML*, 111, at 116.

153 Doxa, as a form of conventional wisdom or a realm of 'truisms' (but not capital 'T' truth) has been correlated with common sense, as a baseline knowledge derived from common experience. See e.g. Georges Molinié, "Doxa et légitimité" (2008) 2 *Langages*, 69. Pietsch also, evoked common intuitions about causality, referring to causal mechanisms thought to be relatively well understood and unambiguous. See Wolfgang Pietsch, *On the Epistemology of Data Science*, *cit. supra*, at 127.

154 Elsa Bernard, *La spécificité du standard juridique en droit communautaire* (Bruylant, 2010), at 37.

155 Laura Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference from Explanation," *cit supra*.

156 H.L. A. Hart, Tony Honoré, *Causation in the Law* (OUP, 1985).

"when two causes are each necessary for producing a certain outcome (conjunctive structure), people judge the abnormal event as more causal."[157]

What is 'normal' and 'abnormal' in the context of AI is open for debate. As we will argue further, the EU's substantive and procedural regulation of AI refers to 'normalcy' by using expressions like 'reasonable foreseeability,' 'intended purpose' (of an AI system), 'foreseeable use (of an AI system) etc. May it suffice stressing, at this stage, that in searching for 'the normal' in connection to AI, scholars' and regulators' reflex was not to focus on descriptive normalcy, but to explore the tenants of a 'new' prescriptive or axiological normalcy. In this context, a 'normally functioning' AI would be one whose output would comply with a given community's foundational axiological framework.[158] In AI jargon, value-conformity is a component of AI accuracy: AI output is 'correct' if it is both statistically accurate (efficacious) and compliant with values labelled as unwavering or norm-setting (effective).

As a flourishing AI scholarship confirms, this stats-meet-values approach to AI accuracy is not the least bit surprising: "new technologies and new forms of human action are always creating moral dilemmas which didn't exist before, which force us to make judgments about how such rules as 'do no harm' apply, and how we interpret or apply the rule in any novel case can only be determined by values external to our rule, values which our rule is in principle incapable of embodying unambiguously."[159] Values,[160] Badea and Artus argue "should be explicit and efficacious, that is, be directly present in the agent's reasoning, and have a material impact upon the decision making of an agent in any relevant situation it acts in. We could then have the agent prioritize these moral goals over practical goals, ensuring that the former are not overruled by the latter."[161] In light of this, the authors suggest that "we adjust the causal power we build into an agent in the design process to the amount which we believe our reasoning mechanisms can successfully handle."[162] If only it were that simple…

AI explainability (and the possibility thereof) are a tricky matter which we will discuss at a later stage in this paper. At this juncture, and after having explored - albeit in broad brush strokes - the objectivist and subjectivist views on explanatory accuracy, a few observations should be made on the importance of explanatory contexts. Indeed, to deliver *good* explanations, explainers should be aware of the intellectual and

---

157 Laura Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference from Explanation," *cit. supra*, at 1489.

158 Axiology is a (vast) field of study with various currents and views on what values are. For the purpose of this paper, the operative understanding of 'value' will be that suggested by Brey, who argued that values correspond to "idealized qualities or conditions in the world that people find good." See Philip Brey, "Values in technology and disclosive computer ethics," *in* Luciano Floridi, *The Cambridge Handbook of Information and Computer Ethics* (CUP, 2012), 41-58, at 46.

159 Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents" (2022) XXXIX International Conference on Innovative Techniques and Applications of Artificial Intelligence (SGAI-AI), 124, at 127.

160 Badea and Artus defined values as "high-level concepts that are relevant considerations during decision making. These could be virtues, character traits ('honesty'), or concepts that are of moral importance ('property') or even morally neutral practical considerations. We argue that values are the tether to the external point of the game, crystallizing what we want from the behaviour of the agents in the game, or in the moral situation. This is supported by arguments from Virtue Ethics." See Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents," *cit. supra*, at 135.

161 *Id*., at 133.

162 *Ibid*.

axiological space in which the explainee operates. The full expression of the '*know thy audience*' rule is in fact, '*know thy audience - in the context where the explanation is received,'* 'context' being understood as the realm of possible experiences which can occur when favorable factors are present.[163]

Why is context so important for explanatory accuracy? Several reasons can be highlighted: because it includes the object of the explanation (facticity); it informs of the explainees' 'background'/conventional knowledge (doxa); it contains the values the explainees look to when deciding if they should believe or not. Above all, context justifies the *inquiries* explainers are called to address. *Explanatory relevance* (the 'why' of an explanation) dictates explanatory *salience* (the 'what' of an explanation), meaning that the answers explanations provide should somehow be *meaningful* in connection to a purpose or interest of importance for the explainee.[164]

To refer back to the example of automated gender bias: the issue of 'how did a system become biased?' naturally calls for informed knowledge (of the system's functionalities) and a capacity to deliver that knowledge (to the satisfaction of the explainee). To provide a 'good answer,' the explainer should exercise so-called *explanatory virtue* which Steel says is "a proxy for probability."[165] The explanatory criteria they should meet are thought to include "the extent to which the hypothesis explains more and different kinds of evidence (*consilience*); the *simplicity* of the explanation, understood as measuring the number and kind of assumptions underpinning it; the extent to which the hypothesis coheres with background beliefs, and the extent to which the hypothesis is *ad hoc*."[166]

Against the backdrop of those criteria, the explainee plays the role of assessor, evaluating whether the explanation given is the *best possible one*.[167] This evaluation essentially takes into account the context in which the explanation is given, the trustworthiness of the explainer and the nature and value of the evidence they bring forward - all factors that may (or not) support the explainee's belief that the information given is *reliable*[168] to a point where it can be seen as accurate, believable or acceptable.

Our general - and for lack of space, lacunary - overview of the epistemology of explanations sets the stage for analyzing this concept's translation in *legal liability* contexts. In those, explanations appear as *instrumental concepts* (means to an end).

---

[163] For an analysis of Boolean probability, in connection to context, see P.D. Bruza, L. Fell, P. Hoyte, S. Dehdashti, A. Obeid, A. Gibson, C. Moreira "Contextuality and context-sensitivity in probabilistic models of cognition" (2023) 140 *Cognitive Psy'y*, 101529.

[164] John Greco, "A (different) virtue epistemology" (2012) 1 *Phil'y & Phenomenological Res.*, 1, at 9.

[165] Sandy Steel, *Proof of Causation in Tort Law* (CUP, 2015), at 79.

[166] *Ibid* (emphasis added).

[167] This echoes the so-called 'best evidence rule', famously coined by Morgan who stated that "the highest degree of probability must govern [courts'] judgment; and it necessarily follows, that they ought to have before them *the best evidence of which the nature of the case will admit*." John Morgan, *Essays upon the Law of Evidence, New Trials, Special Verdicts, Trials at Bar and Repleaders* (Johnson, vol. 1, 1779), at 2-3 (emphasis added).

[168] Steel ties reliability with frequentist probability, the gist being the following: if an explantation stems from frequent occurrences (or causal structures), it will likely be viewed as plausible. See Sandy Steel, *Proof of Causation in Tort Law*, *cit. supra*, at 65: "the evidential probability that p should, plausibly, be influenced by the relevant frequentist probability and (in appropriate contexts) the classical probability that p. The case for p is stronger if there is a very high frequentist probability that p. The point made earlier was only that it cannot be reduced to these."

They are not meant to deliver understanding for understanding's sake; they deliver understanding for the purpose of reaching a verdict, appearing as a crucial component in the exercise of a key public function: the administration of justice.

B. The Accuracy of Causal Explanations

In his study of epistemology in data science, Pietsch raised the question of the function of causal knowledge. Why is it important, he asks, "to identify a relationship as causal rather than as a mere correlation?"[169] The author lamented how misguided we are in believing that *knowing* the causal story is equivalent with being able to explain it: "allegedly, without causal knowledge, one can merely describe how things are, but one cannot explain *why* they are as they are."[170]

The scholar made a criticism and gave a hint. He criticized the 'fundamental mistake' of often confusing causation and correlation: we tend to equate causation with theoretical explanation, while overlooking the much more important function of causation to establish *reliable prediction* and *effective intervention*.[171] Prediction and intervention are, according to Pietsch, what causal knowledge is about:[172] *to know* of a harm-causing fact or event is *to know* how to prevent that fact or event from materializing. Forewarned is forearmed!

Pietsch's hint is one already discussed: as imperfect as they may appear compared to the ideal of scientific knowledge, explanations are, nevertheless, a species of the knowledge-genus. As such, *causal* explanations in law do not translate to an exercise in creative narration but unfold in legally defined procedural frameworks, specifically designed to support reasoning about facts (and the causal links they harbor).

Since explanations deliver understanding (as opposed to 'knowing'), the big question in connection to causal explanations is: what does a court *expect to understand* from an explanation on causation? To answer this question, we will use a distinction, suggested by Le Morvan,[173] between 'knowledge of' and 'knowledge that.' The former is propositional, positing that something is true (e.g. the Earth is a sphere), until proven otherwise. The latter is justificatory, referring to the reasons that justify holding a proposition as true (e.g. there is evidence showing that the Earth is a sphere). Le Morvan's knowledge of/that dichotomy is a useful methodological tool to explore two aspects of causality in law: first, the ways in which causality is *represented* (the 'knowledge of' dimension, explored in **Sub-Section 2.2.1.**); second, the ways in which causality is explained under legally defined standards (the 'knowledge that' dimension, explored in **Sub-Section 2.2.2.**).

---

169 Wolfgang Pietsch, *On the Epistemology of Data Science* (Springer, 2022), at 110.
170 *Ibid* (emphasis added).
171 *Id*., at 112.
172 *Id*., at 111: "(causal knowledge is indispensable) not only for effective intervention but also for reliable prediction. In the absence of a causal connection between different variables, including especially the absence of an indirect connection via common causes, any existing correlation between those variables, no matter how strong, cannot establish reliable prediction."
173 Pierre Le Morvan, "On the ignorance, knowledge, and nature of propositions" (2015) 192 *Synthèse*, 3647.

1. Causality Represented Ex Ante (the 'Understanding of')

As a question of fact, legal causality is first and foremost an issue of evidence. The nature and probative value of the evidence given to support an explanation on causality will, to a large extent, allow a court to distinguish the correlative from the causal, the merely 'possible' from the 'probable.'[174] Think of Spinoza's falling stone. There are at least two plausible explanations for the fall, 1. the wind tilted the stone; 2. God willed the stone into falling. Of course, the evidence supporting each explanation will neither be equally available, nor equally probative: it might be within an inquirer's reach to measure the wind's speed, but it will be far more challenging to elucidate the divine intention behind matters like life-threatening falling objects.

Like in most explanatory contexts, in law, causal explanations are not the products of guesswork or wishful thinking; as factive statements, they need to be backed by evidence, the assumption being that the evidence is, indeed, within the explainor's reach (**A**). However, even when this is the case and evidence is within reach, error is still possible regarding the ways in which causal explanations are given. Two risks in particular are noteworthy: causal *underdetermination* (translating to a narrow view of the causes underlying certain effects) and causal *overdetermination* (translating to a much too broad view of cause-effect interrelationships) (**B**).

a. Da Mihi Facti:[175] the Causal Links Revealed by 'Bare' Facts

Causes - Pietsch writes - can serve as "answers to why-questions even though such answers often do not yield deeper explanations."[176] For example, a layperson might no longer have a headache after taking an aspirin, though they could give only a superficial explanation as to why aspirin cures headaches. Deeper explanations "generally refer either to unifying theoretical laws or to causal mechanisms linking the circumstances with the phenomenon."[177]

Pietsch's view of explanatory superficiality is understandable. There are marked differences in the requirements on 'how far should the discovery of facts go' to meet the standards of, respectively, explanatory and scientific accuracy. The reasons for these differences were outlined in the previous Section. At present, we will focus on the features of the standard of fact-accuracy *required by law*: how 'deep' should the knowledge of causal phenomena be for an explanation thereof to allow the reaching of a fair verdict?

---

[174] By employing the terms 'probable' and 'provable,' we in fact allude to an inductivist theory of legal probability pioneered by L.J. Cohen. Astutely observing (and demonstrating) the occasional absurdity of mathematically calculating the truth value of legal evidence (of innocence or guilt) - as if evidential truth was a measurable property - Cohen suggested a method of *inductive probability*, which departs from an empirical foundation, but is nonadditive and therefore not measurable. This (more 'organic') method of fact assessment is arguably closer to how courts already reason about facts, the example being that of inductive (generalizable) conclusions made based on circumstantial (probabilistically 'weak') evidence. See L. Jonathan Cohen, *The Probable and the Provable* (OUP, 1977).

[175] This adage, in its complete version, is *da mihi facti, dabo tibi jus* - give me the facts and I will give you the law.

[176] Wolfgang Pietsch, *On the Epistemology of Data Science*, *cit. supra*, at 111.

[177] *Ibid*.

The evidentiary and explanatory depth required by law varies, depending on the complexity of the causal constellations the law is called to address. Two uncontroversial statements can be made in this regard. First, causal explanations are usually required *in the presence of harm* having resulted from the violation of a preestablished (typically, legally prescribed) duty of care. Second, the explanations required for the purpose of compensating a harm seek to causally link a conduct, or the reasons underlying it with the harm suffered. As a general rule of thumb, the greater the distance - as it were - between a harmful act and a harm, the greater the 'depth' of the fact-digging enterprise aimed at uncovering the causal chain between the two. While probing evidence is necessary when any causal explanation of a harm is given, its importance is arguably greater in cases of AI-related harm because, in those, discerning the actual causal link is often evidentially challenging in the sense that it is not *directly knowable*. AI use can appear as a 'conduct' having instantiated a harm. However, to understand if a human was involved in that instantiation, it is necessary to understand how the AI system functioned *specifically when the harm occurred* (therefore not generally). In other words, in causal contexts where human and non-human intelligence appear as *plausible candidate-causes of harm*, there is a need for a more in-depth discovery and understanding of the relevant facts.

What does standard liability scholarship tell us about the features of causal explanations? In their seminal work on liability, Hart and Honoré point to two types of causal problems: *explanatory* and *hypothetical*.[178] The former - they argue - "arises when it is not clear how certain harm came about or for what reasons a person did a certain act."[179] The latter arises "when a court, in order to determine whether a wrongful act was in the appropriate sense a necessary condition of the harm inquires whether compliance with the law would have averted the harm."[180] Both deal with the issue of cause, as a precondition for the proof and explanation of causation. Indeed, most debates and evidence in liability cases revolve around uncovering *the* (f)act that can be positively and decisively associated with a harm.

It goes without saying that the concept of cause is relational. Facts - Moore says - are "causal relata"[181] but an *isolated* fact has no causal power. It becomes a cause when, in relation to other facts, it leads to a specific consequence. Hart and Honoré observe that in legal language, the cause-effect dyad is often expressed as 'due to', 'owing to', 'result', 'attributable to', 'the consequence of', 'caused by.'[182] For some purposes - they say - it is important to distinguish between these expressions, "though their similarity on many vital points justifies grouping them together as examples of causal terminology (…) sometimes liability or its extent depends on the proof that a wrongful action, or some other contingency, was the cause of harm: this may be so even where common sense, left to itself, might wish to describe the situation by saying that there were several causes of the harm so each was only a cause."[183] From the perspective of evidence, facts offered as proof of causation seek to establish that an act was indeed 'wrongful' *precisely* because it produced a morally or legally reprehensible

178 H. L. A. Hart, Tony Honoré, *Causation in the Law*, *cit. supra*, at 407.
179 *Ibid*.
180 *Ibid* (emphasis added).
181 Michael S. Moore, *Causation and Responsibility : An Essay in Law, Morals, and Metaphysics* (OUP, 2009), at 33.
182 *Id*., at 87.
183 *Ibid*.

consequence. The important question is - again - that of the criteria used to qualify conduct as 'wrongful' (that is, causally necessary for harm to occur).

The most straight-forward scenario of wrongfulness is that of *unlawfulness*, as exemplified through *legal labelling*. Hart and Honoré alluded to this when referring to hypothetical evidence of causation,[184] the goal of which is to establish that a harm had occurred *because* a legally prescribed duty had not been complied with.[185] However useful, legally prescribed causation can be criticized on two points. First, it tends to synonymize *wrongful* and *unlawful* conduct: harm-causing acts tend to be wrongful, regardless of whether there is a legal rule to confirm that they are. Manslaughter would still be morally wrong, even in the absence of a legal rule to confirm that it was. Alternatively, not all unlawful conduct warrants compensation. Suppose a person got a speeding ticket or was not covered by mandatory health insurance: both are unlawful acts but none creates a duty of compensation, in the sense of liability law. 'Wrongful' acts are therefore a generic category of causal relata which include, but are not limited to 'unlawful' acts, also because no legislator is providential to a point where they can lay out an map of all possible real-world causes and their harmful, compensation-worthy consequences.

This brings us to the second criticism mentioned above: causes (and causations) are vague concepts precisely because no one can have full knowledge of all causal phenomena. Save in rare cases, it is often difficult to *a priori* predict that a specific act has the potential of causing a specific harm. For a swears-by-the-code lawyer, it must be anxiogenic to view the world as an ocean of mostly unforeseeable causal mechanisms which is why law, with its manifest *penchant* for stability, aspires toward *causal invariance*.

b. The Risk of (Mis)representing Causality

To 'represent' or exemplify causality is to have a starting point, a template, an intuition on relevant and repetitive causal connections. However useful, legally exemplifying causality calls for a cautious approach: the 'right' causes should be linked to the 'right' effects. The caution is noteworthy because, as mentioned earlier, reality is causally complex: a cause can have several effects, several causes can converge into producing a single effect, an effect can itself be the cause to some other effect… Causal knowledge is therefore an issue of properly connecting or *fitting together* two or several events, the two obvious risks being that of *underfitting* (tying a cause to one specific effect or set of effects) and *overfitting* (where everything can be the potential cause of everything else). *Adequate* causal knowledge no doubt lies midway between casual underdetermination (***a***) and causal overdetermination (***b***).

i. Causal Underdetermination

Causal invariance is a typical example of underdetermination. It presents itself as an "indispensable navigation device within the infinite space of causal

---

184 H. L. A. Hart, Tony Honoré, *Causation in the Law*, *cit. supra*, at 407.

185 *Id*., at 413: "in the absence of reliable evidence about the hypothetical course of events, a court is naturally inclined to *give effect to the policy enjoining the precaution by assuming*, unless there is evidence to the contrary, that the precaution would have averted the harm" (emphasis added).

representations."[186] It brings reassurance in the face of at least three unpredictable contexts: 1. some important real-world causal interrelationships are unobservable; 2. the environment has the potential to contain unknown background causes of an outcome; 3. it is always possible for background causes of an outcome to differ across contexts.[187]

When there is ambient uncertainty, it is all the more necessary to explain why, say, a specific harm occurred, when the evidence linking it to a cause is not available. Law may then step in and save the day by declaring 'what is what.' This is usually done through *generalizing causal invariance*. When an instrument like the AI Act states that biometric identification systems *typically* cause ethnic discrimination, it generalizes or exemplifies a causal link. This means that all biometric identification systems, past and present, have the potential of developing an ethnic bias which is, of course, an overstretch: they may be perfectly bias-neutral or develop biases on grounds like gender. Causal generalizations are logically 'thin': they take a plausible but narrow belief about reality and convert it into a general, supported-by-the-law example of how reality causally works.

Law has often been accused of being under-deterministic because it tends to introduce simplicity where simplicity is not warranted. An overview of the EU legislation on AI certainly reveals a tendency toward causal underdetermination. As we have argued in a recent study,[188] there is no evidence to overwhelmingly show that biometric identification systems are, *without a doubt* - what the AI Act calls - high-risk systems. On the contrary, we showed that, instead of being evidence-based regulation, the AI Act is primarily a market regulating one, barely relying on facts and mostly giving expression to a seductive value discourse according to which the four levels of risk mentioned[189] are justified by the aim to protect fundamental rights.[190]

In the field of epidemiological evidence, Haack also commented on the not so uncommon disconnection between law and reality: "there can be hard-and-fast rules for determining when epidemiological evidence indicates causation, *the legal penchant for convenient checklists* has led many to construe his list of (…) 'viewpoints' as criteria for the reliability of causation testimony."[191]

Law's causal invariance is convenient but sometimes insufficient because by *labelling causality* it limits the possibility of properly *discovering causality*: biometric identification systems do not develop ethnic biases simply because they perform biometric identification. It is because they - somehow - causally link ethnicity (or any other protected characteristic for that matter) with the purpose for which those systems

---

186 Jooyong Park, Shannon McGillivray, Jeffrey K. Bye, Patricia W. Cheng, "Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions" (2022) 132 *Cognitive psychology*, 1, at 3.

187 *Ibid*.

188 Ljupcho Grozdanovski, Jérôme de Cooman, "Forget the Facts, Aim for the Rights! On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union" *cit. supra*.

189 The four levels of risk in the AI Act are presented *supra* in the Introduction of this paper.

190 See Ljupcho Grozdanovski, "The ontological congruency in the EU's data protection and data processing legislation: the (formally) risk-based and (actually) value/rights-oriented method of regulation in the AI Act" *in* Marton Varju (ed.) *Artificial Intelligence and Law: Values, Rights and Regulation in the European Legal Space* (Springer, 2025), 25 p. (forthcoming).

191 Susan Haack, "Correlation and causation. The 'Bradford Hill criteria' in epidemiological, legal and epistemological perspective," *in* Miguel Martin-Casals, Diego M. Papayannis (eds.), *Uncertain Causation in Tort Law* (CUP, 2015), 176, at 180 (emphasis added).

are used (say, selection of asylum applicants or prevention of crime). *That* causality needs to be uncovered through evidence, even if the evidence reveals causal links other than those that the law (like the AI Act) assigns to specific intelligent systems.

This being said, the discovery of actual, as opposed to preemptively exemplified causation is also tricky because it may show that a harm (say, an unfair bias) can be caused by a plethora of facts or events, each being a plausible candidate to qualify as cause.

ii. Causal Overdetermination

Contrasting law's underdetermination, empiricism faces the risk of *overdetermination.*[192] Pietsch illustrates this with the following example: "the current position of Jupiter might be used by a psychic to scare some poor person to an extent that she commits suicide confirming the very astrological prediction. It seems to follow that the position of Jupiter has to be held fix to fulfill homogeneity when examining causes of suicides."[193]

While courts seldom explain causation in reference to the movement of heavenly bodies, they are not immune to overdetermination. In trials, the risk of overdetermining can occur in essentially two series of cases. First, cases of so-called *concurrent causes i.e*. causes which occur simultaneously and present the equivalent potential of being 'necessary conditions' for a given harm.[194] Second, there is the so-called *pre-emptive kind of overdetermination* where the putative causes are chronologically ordered.[195] Suppose - Moore writes - a building caught fire, and by the time a second fire started, the building has already burnt down.[196] In such a case, we could intuitively assert that the 'necessary' condition for the harm (the burnt building) is the first fire. And yet, a strict counterfactual analysis may yield a "counterintuitive implication that *neither fire* caused the harm because neither fire was necessary (each being sufficient) for the harm."[197] Indeed, with preemptive determination, the problem is that of pinpointing the cause which appears to be the decisive one, in the presence of two or more chronologically ordered or concomitant causal candidates.

The business of linking an effect to its *actual* cause calls for caution in the criteria used to distinguish correlation from causation. This is an issue of both discovery (as an act of evidence-gathering) and explanation (as an act of interpreting the evidence gathered). It is an issue of discovery because the designation of a cause is - here again - largely dictated by the nature and probative value of the items of evidence available. It is an issue of explanation because the evidence is analyzed under specific criteria

---

192 Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, *cit. supra*, at 86.

193 Wolfgang Pietsch, *On the Epistemology of Data Science*, *cit. supra*, at 129.

194 As an illustration of concurrent causes, Moore gives the following example: "two fires, two shotgun blasts, two noisy motorcycles, are each sufficient to burn, kill or scare some victim. The defendant is responsible not for only one fire, shot or motorcycle. Yet his fire, shot or noise joins the other one, and both simultaneously cause their various harms. On the counterfactual analysis, the defendant's fire, shot or noise was not the cause of any harm because it was not necessary to the production of the harms – after all, the other fire, shot or noise was by itself sufficient." See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, *cit. supra*, at 86.

195 *Ibid*.

196 *Ibid*.

197 *Ibid* (emphasis added).

used to determine if one cause or a chain of causes had, indeed, a *decisive* influence on the harm materializing. In complex causal scenarios - where the cause-harm link is not straightforwardly discernable - the decisiveness aspect is usually uncovered through the search for the so-called *proximate cause*. Hart and Honoré tell us that under the heading of 'proximate cause' we find multiple methods of causal fact assessment. Almost always - they say - the relevant question is whether or not the harm would have happened without the defendant's act: "this factual component is variously termed 'cause in fact' 'material cause,' *conditio sine qua non*, and is the sole point of contact with what causation means apart from the law."[198] The authors further explain that the 'proximate' label can be used to explain (or not) cause-effect occurrences and is often given out of convenience, public policy, "a rough sense of justice"[199]… When a law decides that beyond a certain threshold of probability, the cause is no longer 'proximate' (*i.e*. can no longer be positively associated with a harm) causation becomes an issue of "practical politics."[200]

Proving and assessing the *degree of proximity* between a possible cause and a harm becomes even more complex when an alleged harm-doer appears to be causally far removed from the harm. This is usually the case in so-called *material contribution* cases and *vicarious liability* cases. In the former, the victim is typically required to show that the defendant's wrongful conduct had made "a 'material contribution' to the disease or injury. The doctrine of material contribution applies to conditions (…) which are known often to be caused by prolonged exposure to some agent (e.g. dust) but where the effect of any particular period of exposure is hard to argue."[201] Material contribution is a *faute de mieux* approach to causation, typically "when the actual cause of an occurrence is unknown in the sense that there is not sufficient evidence to show in detail what happened on the occasion in question."[202] In such a case, a court would look for evidence of "the characteristically different processes by which different causes produce their effects.[203]

In vicarious liability cases - relevant for commodities such as AI - the causal connection sought is that between "the servant's action or omission and the harm, and in no sense of causation is it necessary to establish any causal connection between the master's conduct and the harm."[204] In a scenario involving AI, the 'servant' would be the artificial system whose decision or prediction would act as the *apparent cause* of harm. However, the 'master' would - always - be a human agent exercising a legal right (ownership, use) and complying with a duty (e.g. control and oversight) over that system. The issue of AI liability will be discussed in more detail further.[205] At this juncture, may it suffice stressing that the world of causation (and the explanations thereof) is rich and complex, lending itself to a variety of explanatory possibilities. Let us, therefore, bring forward accuracy as *fil rouge* of this Section.

With accuracy in mind, the legally relevant issue becomes the following: *in a specific case* (*ergo* not generally), how can a harm be plausibly, if not accurately,

198 H.L.A. Hart, Tony Honoré, *Causation in the law*, *cit. supra*, at 90.
199 *Ibid*.
200 *Ibid*.
201 *Id*., at 410 (emphasis added).
202 *Ibid*.
203 *Ibid*.
204 *Id*., at 85.
205 See *infra*, Sub-Section 4.3.

viewed as the consequence of a conduct, phenomenon or event? This is not an issue of *deontic reasoning* (what the law orders us to view as cause). It is an issue of *practical reasoning* based on (and presupposing) a source of valid, trustworthy empirical information that supports the understanding of the relevant facts, offering an answer to a causal inquiry (e.g. who or what *actually* caused a harm?).[206]

In short, causation is a matter of getting the *right kind of evidence* and delivering the *right kind of understanding* based on that evidence because - as will be argued in the following sub-section - in the presence of multiple candidate-causes, *justice* requires that the actual cause of a harm be uncovered. In other words, what we're aiming at is distilling causation from a sea of correlations.

2. Causality Explained Ex Post (the 'Understanding That')

As argued previously, to explain causality is to give an 'accurate' (believable[207]) account of the various stages of a causal chain that connect a fact with an end-result (typically, a harm). We also alluded to the fact that the problem with AI is that the opacity of automated decisional processes makes it difficult to straightforwardly establish a cause-effect connection. Indeed, direct and probing evidence in support of causal explanations is often unavailable, pushing courts to call for expertise which - as the caselaw shows - may neither be available, nor clear on how a well-performing system should and is likely to operate (**A**).

If and when evidence on possible cause/effect correlations is given, courts typically seek to separate causal from correlative associations. To do so, liability doctrines and court practice offer a series of so-called causality tests: essentially, forms of counterfactual reasoning designed to determine if a fact, event or trope was both sufficient and necessary to yield a specific harmful result (**B**).

a. Lessons from North American Caselaw in the Field of AI Liability

The available examples of judicial instances in AI liability - mostly brought before North-American courts - give valuable insight into the evidence that both litigants and courts flag as necessary and probative for the purpose of explaining causation in connection to 'harmful' AI systems. To induce conclusions - as useful takeaways for the future application of the EU's regulation on AI liability - we will focus on the two, abovementioned set of factors that impact explanatory 'goodness.'

On the one hand, we argued that explanations are fact- and context-bound, their 'goodness' being largely dictated by the evidence of the facts that fall in the scope of the explanations. In the existing AI liability caselaw, *expertise* emerges as a privileged mode of evidence (***i***). On the other hand, we argued that a 'good' explanation is one that warrants believability: a situation where the explainee considers they have sufficient reason to accept an explanation as plausibly true. In the caselaw cited hereafter, two trends emerge regarding the conditions for believability litigants and

---

[206] Friedman rightly pointed out that "if epistemic rationality is a form of instrumental rationality, following one's evidence should be conducive to achieving one's epistemic goals." Jane Friedman, "Teleological epistemology" (2019) 176 *Phil. Studies*, 673, at 677.

[207] We allude to our comments on believability as standard for explanatory accuracy, see *supra* , Sub-Section 2.1.2.

courts appear to observe, when assessing if an explanation on a harmful AI system warrants acceptance (***ii***).

i. Lessons on the Fact-Correspondence of Causal Explanations: Expertise as a Preferred Type of Evidence

To distil causation proper from a multitude of correlations, causal explanations - factive as they are - require tangible, probing and verifiable evidence of the causal link between a defective product (like a biased AI) and a harm suffered (say, gender discrimination). When *direct evidence*[208] of that link is unavailable, courts may turn to expertise, the admissibility of which is usually framed by procedural requirements of 'scientificity,' reliability and trustworthiness.

In the US, Rule 702 of the Federal Rules of Evidence defines the essential features that expert evidence must present to be declared admissible. This provision states that "to be scientific knowledge (…) valid reasoning and methodology must be employed: (1) peer review and publication, (2) the known of potential rate of error, (3) general acceptance and (4) testing a theory by attempting to find evidence to disprove it ('falsification')."[209] The main purpose of these criteria is to support the monitoring of the reliability of expert testimony, allowing courts to 'weed out' so-called 'junk science.'''[210]

The criteria listed in Rule 702 are, in a sense, a codification of the 'original' expertise case *i.e. Frye.*[211] In this case, a person was being tried for murder. In their defense, they called an expert witness who testified on the results of a systolic blood pressure deception test, the argument of the defense being that blood pressure was influenced by the changes in the witness's emotions, being on the rise when the witness experienced nervousness. The obvious issue here was whether such a test could be admitted as legal evidence. The court's approach on this point was cautious: while it did not altogether dismiss scientific expertise as a mode of evidence, it defined a key admissibility requirement which referred to the *epistemic soundness* of the method used to yield a result the court might decide to consider as probing. Since judges are not scientists, the criterion used to determine if a method of discovery produced valid knowledge (as opposed to speculative information), it was stated in *Frye* that "while the courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to *have gained general acceptance* in the

---

208 In evidence scholarship, direct evidence (as opposed to indirect evidence) is usually understood to mean proof of fact which does not call for the reality of that fact to be inferred. According to Cansacchi, the 'directness' of evidence derives from the type of information an item of evidence reveals to an assessor (say a court). Direct evidence brings a reality directly to the knowledge of the assessor, without requiring any mediation (additional items of evidence) and without inviting the assessor to interpret what the item can mean. See Giorgio Cansacchi, *Le prezunzioni nel diritto internazionale : contributo allo studio della prova nel processo internazionale* (Eugenio Jovene, 1939), at 11.

209 Michael D. Green, Joseph Sanders, "Admissibility versus sufficiency. Controlling the quality of expert witness testimony in the United States," *in* Miguel Martin-Casales, Diego M. Papayanis (eds.), *Uncertain Causation in Tort Law* (CUP, 2015), 203, at 214.

210 *Id*., at 204.

211 Court of Appeals of the District of Columbia, 3 December 1923, *Frye v. US*, 293 F. 1013 (D.C. Cir. 1923).

particular field in which it belongs."[212] We find here a procedural translation of the principles of acceptance of knowledge discussed earlier:[213] a scientific community is likely to 'validate' knowledge based on the soundness and reliability of the methods used to produce it.

Since *Frye* (1923), the conditions under which the 'general acceptance of a scientific method' could be declared were further clarified in *Daubert*.[214] In this case, the parents of two minor children with birth defects alleged that those defects were due to the mothers' prenatal ingestion of a prescription drug marketed by the defendant. The probative issue was whether the available expertise revealed a risk that the drug might indeed be causally linked to those defects (which experts largely denied). The merit of *Daubert* is that it provides useful insight into the criteria applied to determine the probative quality of scientific expertise. Those criteria pertain to the *trustworthiness* and *admissibility* of expertise and to its *impact* on the outcome of a dispute.

On the point of trustworthiness and based on both *Frye* and Rule 702 of the Federal Rules of Evidence, the Supreme Court in *Daubert* first formulated the basic accuracy requirements, specifying that the adjective 'scientific' implies a "grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation."[215]

Within the framework of our discussion of objectivism/subjectivism in connection to scientific knowledge,[216] the US Supreme Court is - understandably - subjectivism-averse, since probative 'knowledge' cannot be reduced to mere 'subjective beliefs.' The Supreme court further distinguished between validity and reliability, although "the difference between accuracy, validity, and reliability may be such that each is distinct from the other by no more than a hen's kick."[217]

Translating this validity/reliability distinction in the context of dispute-resolution, the Supreme Court noted that "our reference here is to evidentiary reliability that is, trustworthiness."[218] In the interest of assessing the level of general acceptance of a discovery method, a "reliability assessment does not require, although it does permit, explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community."[219] *Widespread acceptance* "can be an important factor in ruling particular evidence admissible"[220] whereas "a known technique which has been able to attract only minimal support within the community (…) may properly be viewed with skepticism."[221] The focus - the Court stated - should be "*solely on principles and methodology, not on the conclusions that they generate*."[222] Based on these premises, the Court's conclusion was obvious: expert knowledge given as evidence in a trial

---

212 *Id*., at 1014 (emphasis added).
213 See *supra*, Sub-Section 2.1.2.
214 US Supreme Court, *Daubert v. Merrell Dow Pharmaceuticals*, 28 June 1993, 509 U.S. 579 (1993).
215 *Id*., at 590.
216 See *supra*, Sub-Section 2.1.
217 US Supreme Court, *Daubert v. Merrell Dow Pharmaceuticals*, *cit. supra*, at 590.
218 *Id*., at 592.
219 *Id*., at 594.
220 *Id*., at 595.
221 *Ibid*.
222 *Ibid* (emphasis added).

should meet at least basic validity requirements warranting acceptance in the relevant scientific field.

More interestingly on the second point - pertaining to the expertise/fairness interrelationship - the Supreme Court stressed that 'scientific' evidence, albeit relevant, can be excluded from a trial "*if its probative value is substantially outweighed by the danger of unfair prejudice*, confusion of the issues, or misleading the jury (…) Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it. Because of this risk, the judge in weighing possible prejudice against probative force under Rule 403 of the present rules exercises more control over experts than over lay witnesses."[223]

Regarding fairness, the ruling in *Daubert* is truly eye-opening because if confirms the specific status of *science-based judicial truth*: accuracy of the disputed facts is, indeed, a precondition for an informed, impartial and by that, fair adjudication. However, courts must remain mindful of the *finality of fairness* of adjudicatory procedures. This is especially true in cases - like those analyzed further in this paper - where consensus on a scientific method is not widespread, but the legal stakes of verifying the soundness of that method are high, especially in criminal proceedings where accurate and reliable information is paramount for the issuing of a verdict. Law asks for fairness and expediency while scientific discovery is ever evolving and seldom set in stone: "scientific conclusions are subject to perpetual revision. Law, on the other hand, must resolve disputes finally and quickly."[224]

In this context, general acceptance, as originally defined in *Frye* was to be viewed as "not a necessary precondition to the admissibility of scientific evidence under the Federal Rules of Evidence, but the Rules of Evidence - especially Rule 702 - do assign to the trial judge the task of ensuring that an expert's testimony both *rests on a reliable foundation* and is *relevant to the task at hand*. *Pertinent evidence based on scientifically valid principles will satisfy those demands*."[225]

The moral of the story in *Daubert* is that fact-accuracy is of course important, but it is function of *evidentiary relevance*: scientific expertise, when used in the courtroom, is not meant to answer a question of pure science; it *participates in answering a question of law*. In other words, the admissibility of (scientific) expert evidence should not rely solely on scientists' opinions *en général*; it should meaningfully guide a court in the latter's application of the law to a specific factual situation.

Following *Frye* and *Daubert*, Anglo-American scholarship explored subsequent applications of this caselaw, in an attempt to induce general criteria (or court trends) used to assess the reliability of scientific expertise. Braford-Hill[226] famously suggested a list, arguing that reliability of scientific evidence - especially in causal scenarios - is most frequently function of the *strength* of the causal association,[227]

---

223 *Ibid* (emphasis added).
224 *Id*., at 597.
225 *Ibid*.
226 See Austin Bradford Hill, "The Environment and Disease: Association or Causation?" (1965) 58 *Proceedings of the Royal Society of Medicine*, 295.
227 Susan Haack, "Correlation and causation. The 'Bradford Hill criteria' in epidemiological, legal and epistemological perspective," *cit. supra*, at 182.

consistency (stemming from the converging results from different investigations performed in different places),[228] *specificity* (the association should be restricted to a specific cause-effect interrelationship), [229] *temporal precedence* (the cause must consistently precede the harm)[230] a *gradient* (essentially a threshold of gravity)[231] and *plausibility* (the cause-effect connection should be plausibly considered as causation),[232] *coherence* (the causal interpretation should not seriously conflict with known facts about the cause-effect interrelationship).[233]

The trouble in AI litigation is that expertise, fitting all of the Bradford-Hill criteria, is often not available. More often than not, direct evidence of a system's 'inner workings' - at the time a harm occurred - will not be available. In a world where transparency and explainability would reign, whenever harm would be causally linked to an AI system, an independent expert would be called to reverse-engineer that system's decisional process, zooming in on the point where the harm-causing 'glitch' appeared. However, save in cases of fully transparent and explainable systems, the scenario of experts stepping in to crack open the black box and save the day is not, and will not be as frequent. If independent expertise is not likely to be feasible, which evidence can courts rely on to discern causation? The *Pickett*[234] and *Loomis*[235] cases can shed some light in this regard.

ii. Lessons on the Believability Dimension of Causal Explanations: the Types of Understanding Sought

(1) The Understanding Sought by Courts: the Shift from 'What Experts Prove' to 'What Experts Say' in Pickett

In 2017, two police officers travelled in an unmarked vehicle in New Jersey. A group of men wearing ski masks and armed with handguns fired in a crowd causing the death of one person. Shortly thereafter, they were arrested. A ski mask, recovered by the police, was analyzed for DNA. The analysis showed two specimens of saliva. A buccal swab from the suspects showed that one of them was the main source contributor. The remaining specimen could not be analysed using traditional DNA testing. The samples were then sent to Cybergenetics (a private laboratory), owner of the TrueAllele software program, assumed to be far superior in terms of accuracy to traditional forensic DNA tests, especially when dealing with complex DNA mixtures. The results correlated the DNA specimen to the defendant (Pickett). He challenged the accuracy and reliability of the probabilistic genotyping, calling for *independent studies* to investigate whether TrueAllele correctly applied the probabilistic genotyping methods.

---

228 *Id*., at 183.
229 *Ibid*.
230 *Id*., at 184.
231 *Ibid*.
232 *Ibid*.
233 *Ibid*.
234 Superior Court of New Jersey (Appellate Division), 2 February 2021, *State of New Jersey v. Corey Pickett,* Docket N° A-4207-19T4.
235 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, 881 N.W. 2d 749 (2016) 2016 WI 68.

The experts stressed that the software program contained approximately 170'000 lines of code written in MATLAB (a programming language designed specifically for visualizing and programming numerical algorithms).[236] They claimed it would take hours to decipher a few dozen lines of the 'dense mathematical text' comprising the code,"[237] leading up to "about *eight and a half years* to review the code in its entirety."[238] In other words, reverse-engineering was not an option, namely because of the excessive duration which would adversely impact the reasonable duration of the trial. In the absence of expert evidence, the courts reverted to *alternative evidentiary strategies*. The question having guided their reasoning is the following: if an *expert cannot prove* the accuracy of TrueAllele's decision in the specific case of Pickett, *what do experts say* on the system's aptitude for accuracy *in general*?

This shift from '*what can experts prove*' to '*what do experts say*' has an important procedural repercussion because it shifts the debate from evidence that is case-specific, highly probative but unavailable (reverse-engineering) to information that is available, but not case-specific and not particularly probative (general expert opinions). Following this approach, the Attorney General in *Pickett* considered three types of evidence: the testimony given by Cybergenetics' expert, validation studies and publications on TrueAllele and opinions from other jurisdictions on the system's performance. All three types of evidence converged on the point that TrueAllele was, *in principle*, reliably accurate[239] which, of course, triggered some discontent.

It was argued that *general expert acceptance* of a model's accuracy (providing, at best, presumptive evidence of debatable probative force) is no substitute for independent, unsupervised review of the source code (providing direct evidence, with strong probative force).[240] It was also argued that even simple software programs are "prone to failure, and that an error in any one of the three domains of software engineering - problem identification, algorithm development and software implementation - undermines the trustworthiness of the science underlying the relevant expert testimony."[241] These opinions are, of course, legitimate. But if *Pickett* confirms anything about the quality of evidence used for the purpose of arriving at a (plausible) causal explanations, it is that in future AI liability cases, the most conclusive evidence (reverse-engineering) may not always be within reach. In cases where the *probatio* (expertise on AI concrete performance) is unavailable, courts are likely to turn to *fama* (an AI's reputed performance in a majority of cases). This redirection from *in concreto* evidence assessment to general opinions is what Duede called *brute inductive consideration i.e.* a belief that an AI system is reliable based on past reliability evaluations.[242] And such a reasoning is 'all too human': given that AI systems can be opaque (therefore, inscrutable and unpredictable), courts 'naturally' search for expert

---

236 *Id*., at 17.
237 *Ibid*.
238 *Ibid* (emphasis added).
239 *Id*., at 25.
240 Some experts stated that the reliability of the TrueAllele software "cannot be evaluated without full access to 'executable source code and related documentation,' something that no one to date has seen." See *State of New Jersey v. Corey Pickett*, Docket N° A-4207-19T4 *cit. supra*, at 34.
241 *Id*., at 35.
242 Eamnon Duede, "Instruments, agents, and artificial intelligence: novel epistemic categories of reliability" (2022) 6 *Synthese*, 1, at 3. Audi called this derivative reliability which, in essence, warrants trust in an information based on the reliability of the source of that information. See Robert Audi, "Reliability as a virtue" (2009) 142 *Phil. Studies*, 43, at 46.

opinions that can confirm a system's *behavioral consistency*. But is this good enough from the litigants' perspective? The answer is 'no;' the *Loomis* case gives hints as to why.

(2) The Understanding Sought by Litigants: the Reasons for (Human) Reliance on AI Output in Loomis

Loomis[243] deals with the use of COMPAS, a risk-need assessment tool designed to predict recidivism and to identify program needs in areas such as employment, housing and substance abuse. The claimant was accused of being involved in a drive-by shooting which he denied. He was charged with five counts and pleaded guilty to only two of the less severe charges. After accepting Loomis's plea, the circuit court ordered a presentence investigation which included a COMPAS risk assessment. The risk scores in this assessment were intended to predict the general likelihood that those with a history of offending are either less likely or more likely to commit another crime following their release from custody. The prediction was based on a comparison between information pertaining to an individual and information pertaining to members of a similar data group. It should be stressed that the risk scores produced by COMPAS were not intended to determine the severity of the sentence or whether the offender should have been incarcerated.

In *Loomis*, the defendant contested the court's *reliance* on COMPAS's allegedly biased prediction which resulted in predicting a higher risk of recidivism, naturally leading to a more severe sentence. In essence, the defendant contended that *by slavishly relying on COMPAS*, the sentencing court erroneously exercised its discretion by not basing its decision on other facts in the record. The consequence of this - it was argued - was the violation of the defendant's due process rights namely, the right to be sentenced "based on accurate information;" the right to an individualized sentence and the improper use of gendered assessments in sentencing.[244]

Loomis is foretelling of a caselaw we will likely see develop in the future because it points to the *reasons underlying the human reliance on a given AI output*. Indeed, the evidentiary (and explanatory) issues we will see down the line will likely not only focus on whether the author of harm was an AI or a human, but if it was a human agent's slavish (non-reasoned) reliance on an automated decision/prediction that caused the harm. To make their argument in this sense, a litigant would need to demonstrate that: 1. an AI output was inaccurate (e.g. biased), which would require proof and explanation on the system's functioning and performance; 2. that the reliance on that output was harmful, which would require evidence on *pre*-, *prae*- and *post*-use accuracy checks.

Like in *Pickett*, in *Loomis*, reverse-engineering of COMPAS was not performed. Rather, the Wisconsin Supreme Court turned to sources, external to the dispute, to arrive at a conclusion on the system's *general accuracy* (thus confirming the above-mentioned shift from *probatio* to *fama*). The Court found e.g. that some States - like New York - have conducted validation studies of COMPAS concluding that its risk

---

243 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra*.

244 *Id*., § 34.

assessments were generally accurate.[245] The defendant, however, cited a 2007 California Department of Corrections and Rehabilitation (CDCR) study which concluded that there was "no sound evidence that COMPAS can be rated consistently by different evaluators (…) that it predicts inmates' recidivism."[246] This study notwithstanding, the Wisconsin Supreme Court considered that the sentencing court used COMPAS merely as an 'aid,' not as a basis for its decision. But the 'battle of experts' in *Loomis* is not reassuring because - again - general expert opinion is hardly strong proof of a system's accuracy *in a specific case*. What would have happened if the California study, critical of COMPAS, was seen as more probative than other (contradicting) studies?... The relevant caselaw is too embryonic to infer the criteria used by courts in their selection of reliable and trustworthy expertise, as regards the aptitude for accuracy of AI systems. For the purpose of this paper, we will view *Pickett* and *Loomis* as examples showcasing an emerging (but not consolidated) trend of *casting a wide net on evidentiary relevance*: when concrete, case- and AI-specific expertise is desirable but unfeasible, general (and reliable) expert opinions on the AI concerned will have to do.

The cited cases also illustrate that evidence is but the first step of the causal explanatory enterprise in cases of AI liability. Save in rare instances where evidence is self-explanatory (e.g. the training data reveals the presence of a bias) the items of evidence discussed before a court will usually be integrated into explanatory narratives which - as mentioned earlier - will aim at delivering causal understanding that explainees (*i.e*. courts) can 'buy into.'[247] To assess the level of understandability and believability, courts use a number of so-called causality tests. These usually play an *exclusionary role*: they are meant to allow the assessment of the 'goodness' of the understanding that explanations deliver, in view of eliminating those which (plausibly) show *correlation* from those that (plausibly) show *causation*.

b. The 'Tests' Used to Explain Causation: But-For and its Variants

Causality 'tests' are used in many legal systems but have especially been developed in Anglo-American court practice and statutory evolution. There are usually notable differences in the ways in which they apply, depending on whether causation is proven in the context of tort or criminal law.[248] As a general *summa divisio* - and based on Moore's work - these tests can be perceived as variations of one test, seen as fundamental across Common law systems: the *sine qua non* or *but-for* test.

This test supports the following *counterfactual reasoning*: but for the defendant's action, would the victim have been harmed in the way that law prohibits?[249] In both criminal law and tort law - as well as in direct and proximate cause scenarios - the but-for test allows courts and juries to zoom in on two points which, if supported by evidence, are likely to uncover the causal or correlational nature of a fact/harm link.

---

245 *Id*., § 59.
246 *Id*., § 60.
247 See *supra*, Sub-Section 2.1.2.
248 Moore argues that criminal law has been a 'borrower' from torts regarding the 'tests' aimed at proving and assessing causation. However, this "borrowing has not been uniform and without reservation (…) the criminal sanction of punishment is sometimes said to demand greater stringency of causation than is demanded by the less severe tort sanction of compensation." See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, *cit. supra*, at 83.
249 *Ibid*.

These points are the *necessity* of the cause for the harm to occur (meaning that without a specific event acting as cause, a harm would have not materialized) and the *sufficiency* of that cause (meaning that the cause was a determining factor for the harm to materialize). By applying the counterfactual reasoning based on the but-for test, court practice has developed a series of variants - specific tests to assess causal necessity and sufficiency. Moore cites, as examples, the *necessary element test*; the *necessary to the time, place and manner* of an effect's occurrence; *asymmetrically temporal* test, the *necessary to accelerations* test; the *necessity of negligent aspects* of acts; necessity as a usually present and always sufficient criterion of *'substantial factor' causation* and causation as *necessity to chance*.[250]

In the field of AI, a peculiar application of the but-for test can be detected in *Loomis*.[251] In assessing the sentencing court's reliance on COMPAS when reaching a verdict (as the causal issue in this case), the Wisconsin Supreme Court found that, the COMPAS assessment was not "determinative in deciding whether Loomis should be incarcerated, the severity of the sentence or whether he could be supervised safely and effectively in the community."[252] To support this argument, the Wisconsin Supreme Court applied a peculiar 'but-forian' reasoning, arguing that the circuit court *would have imposed the exact same sentence* even without having used the COMPAS system.[253]

Loomis gives a glimpse into the reasoning courts are likely to apply in future AI liability cases which will depart from the following question: *would the user of the AI system arrive at the same (harmful) decision, had they not used the system in the first place*? Asking this question is tricky because it opens the door to speculation. To avoid this, we will perhaps see the emergence of additional tests down the line. For example, a '*reasonable user*' test might emerge, which would translate to examining an agent's conduct in a specific occurrence and seek to determine if the alleged harm would have nevertheless occurred, without that agent's conduct.[254] It is - again - too early to speculate on the ways in which the but-for test might be applied in future AI liability cases.

The *second type* of tests include a variety of policy-based tests such as the *reasonable foreseeability* and *harm-within-the-risk* tests. According to Moore, the goal of those is to "describe a factual state of affairs that plausibly determines both moral blameworthiness and duties to compensate, and that plausibility connects a defendant's

---

250 *Ibid*.

251 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra*.

252 *Id*., § 109.

253 *Id*., § 110.

254 Bathaee argues that the principal-supervision rule derived from standard principles on agency can applies in assessments of causation in AI liability cases. The test would basically seek to establish if the programmer or user of the AI system exercised reasonable care in processes like monitoring, designing, testing or deploying. Of course, the principal-supervision rule is applicable in instances where supervision is possible. In cases of unsupervised ML, the relevant issue - Bathaee stresses - is whether it is at all reasonable to have used or deployed such a system. The answer, the scholar says, may be no, which would mean that the creator or user of that system would be liable for any harm that it might cause. See Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation" (2018) 2 *Harv. J. L. & Tech'y*, 890, at 936.

culpability to particular harms."[255] These are the tests we alluded to when we discussed causal invariance (where the law connects specific causes to specific harms).[256]

The *harm-within-the-risk* test essentially serves to discern causation when a cause is associated with - so to speak - a *family of harms*.[257] Think of a recruitment AI: though they are commonly associated with gender biases, we would hardly be surprised if they, at some point, expressed an ethnic bias. Before we witnessed the outcome of the EU's rights-matter-to-us regulatory framework on AI (AI Act, AILD), part of scholarship - including the author of this paper - pleaded in favor of an acceptance-of-risk criterion, serving as referent for identifying the agent having accepted that an AI system may cause harm and can, because of that acceptance, be held responsible.[258] The AI Act essentially integrates the *harm-within-the-risk test* by introducing a form of causal invariance for so-called high-risk AI systems. The invariance aspect is visible in the list of sectors and uses that the AI Act flags as falling in the 'high risk' category. For example, in the field of migration, asylum and border control management, it mentions systems used as polygraphs (and similar tools) aimed at detecting emotional state(s) of a natural person. Intuitively, we could agree that this is, indeed, a high-risk use: errors in detecting emotional states can produce unwanted consequences, especially when such detecting is performed in the processing of asylum applications. The procedural question is whether this causal invariance in the AI Act would somehow *lighten* the burden for victims to prove causality. Imagine an asylum seeker who underwent an emotion recognition test which concluded that the applicant was lying when they explained the reasons why they were forced to flee their country of origin. Based on that decision, their asylum application would presumably be rejected. Suppose the applicant wished to contest that rejection. Would they be required to prove the cause (the system's error) and its harmful consequence (the rejection of the asylum application), given the AI Act states that emotion recognition systems are 'high risk' *anyway*? Now that we have the EU's AI Liability framework, the answer is 'no': though the list of 'high risk' systems in Annex III of the AI Act integrates a causal invariance rationale, it does not create a *general presumption of harm and causation* when those systems are used in practice. The AI Act merely circumscribes the scope of the harms associated with 'high risk' AI, but does not include a *general liability test*, nor does it attach any procedural consequence (e.g. discharge of the burden to prove harm) for high-risk systems. The evidentiary issues associated with those systems are addressed in the EU's forthcoming legislation on AI liability, which will be analyzed further in this paper.

Under the *foreseeability test*, the relevant question to ask is whether a harm was intended, foreseen and foreseeable enough "to render any actor unreasonable for not

---

255 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra,* § 110.

256 See *supra*, Sub-Section 2.2.1. (B).

257 Moore writes that "the harm-within-the-risk test is in the service of justice-oriented policy in its seeking of a true desert-determiner and the test does not ask a redundant question (…) The real question for the harm-within-the-risk test is whether the grading by culpable mental states is all that is or should be going on under the rubric 'legal cause.'" See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics, cit. supra*, at 100.

258 Ljupcho Grozdanovski, « L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale ? Réflexions sur l'avenir du régime de responsabilité du fait de produits défectueux dans l'Union européenne », *cit. supra*.

foreseeing it."[259] Of course, the specificity of this test is that it necessarily includes a *subjective element*. Moore calls it the *fit problem*: "fact-finders have to fit to the mental state of the defendant had to the actual result he achieved and ask whether it is close enough for him to be punished for a crime of intent."[260] In criminal law, intent is paramount considering that, for many criminal offences, evidence of the intent-to-harm is required. An interesting example of foreseeability and the (impossible) proof of *mens rea* in the field of AI is given by the *Coscia* case.[261] Here, high-frequency trading algorithm performed spoofing *i.e.* placed phantom orders in the market, then withdrew them when the markets began to 'move' in a desired direction. Since spoofing is a criminal offence in US law, the proof of spoofing requires evidence of *intent to harm* (*mens rea*), which the algorithm in *Coscia* of course did not have. The US courts' found themselves in an unenviable position: on the one hand, they were held to ask for and assess intent-to-harm evidence but were, on the other hand, faced with an objective, practical difficulty to access such evidence, since AI autonomy does not include intentionality *per se*. In this procedural setting, the courts' reflex was to, essentially, *broaden the scope of admissible evidence* and require that the parties '*prove until a responsible human is found*.' Testimonial evidence was ultimately key in adjudicating this case: it was the system's programmers who, in their testimony, revealed that it was the user who 'commissioned' a system capable of spoofing.

In *Coscia*, the intent-to-harm test, when applied, did ultimately direct the court to a human agent. We may however imagine and even expect instances where this might not be the case, leaving open the question of the human who ought to be criminally responsible when *no evidence* shows any trace of criminal (human) intent. This issue will likely not be raised in the EU, since the AILD regulates civil liability. But national courts (including those of the EU Member States) may, at some point in the future, be confronted with scenarios like the one in *Coscia*, only without testimonial evidence to guide them to a responsible human.

II. ACCURACY IN CONNECTION TO EXPLAINABLE AI (XAI)

In connection to AI, accuracy is a tricky concept for two reasons. First, on a theoretical level, AI technologies are slowly pushing changes on some of the bedrock-principles of epistemology: we are now in the era of data-driven science which "seeks to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon."[262] This new field of data science seeks to "incorporate a mode of induction into the research design, though explanation though induction is not the intended end-point (as with empiricist approaches)."[263] Instead, "it forms a new mode of hypothesis generation before a deductive approach is employed. Nor does the process of induction arise from nowhere, but is situated and contextualized within a highly evolved theoretical domain."[264]

---

[259] Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics, cit. supra*, at 100.
[260] *Ibid*.
[261] US Court of Appeals for the 7th Circuit, *US* v. *Coscia*, 866 F.3d 782 (2017).
[262] Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts" (2014) 1-1 *Big Data & Society*, 1, at 5.
[263] *Id*., at 6
[264] *Ibid*.

Second, and more importantly, there is the question of *law's response* to these new 'epistemic actors.' The fundamental issue here is whether evidentiary causal explanations in AI liability cases can, or even should integrate explanations of specific AI output (*i.e*. if causal explanations about AI-related harm require explainable AI).

Before we tackle this issue in the context of the EU's procedural framework on AI liability, we should pay closer attention to the criteria according to which AI output can be viewed as accurate (**Sub-Section 3.1.**). In light of those, we will then explore the conditions that *explanations on AI* should meet in order to, themselves, be qualified as accurate or, at the very least plausible (**Sub-Section 3.2.**).

A. Accuracy Standards for AI Output

When Badea and Artus defined 'intelligence' in connection to *artificial* intelligence, they gave the impression of weighing their words and rightfully so: the only referent we have for intelligence is that of *human* intelligence which smart technologies are capable of simulating, without - yet - fully reaching the intelligent-as-a-human standard: "by intelligence, we of course do not necessarily mean anything as grand as consciousness or Artificial General Intelligence (AGI), but, rather, the *ability to be an effective and creative utility* (or function) maximiser, *i.e*., a machine that is 'clever' at finding ways to achieve the goals we set for it."[265]

But machines can be 'clever' in achieving preassigned goals in ways that humans (clever as they themselves are) are not always capable of discerning or foreseeing. In this context, the question 'what is *accurate* AI output?' depends on first addressing the issue of 'how does AI produce knowledge of the world in the first place?' To address these questions, it is necessary to first explore the peculiar epistemic status of intelligent technologies which albeit created by humans, gradually become their (mighty) fellow-knowers (**Sub-Section 3.1.1.**). Against this backdrop, we can then explore the challenges that humans experience when explaining how AI systems actually 'understand' information about reality (data), when they have nothing else to go by but the output those systems produce (**Sub-Section 3.1.2.**).

1. The Epistemic Specificity of Non-Human 'Knowers'

From the perspective of 'standard' knowledge-construction theory[266] whereby human agents are the sole 'knowers' of the world, AI technologies are certainly avantgarde: for the first time in history, non-human entities are capable of employing the reasoning models historically associated with humans. Because of this, we would be inclined to assume an *epistemic parallelism* between human and non-human 'knowing': since both deploy the same reasoning models, they must also share the same standards by which the knowledge they acquire can qualify as accurate. A nuance should however be highlighted. It is one thing to draw parallels between humans and AI on how they go about acquiring knowledge. It is another thing to inquire on how humans arrive at such knowledge when the object they seek to 'know' (or understand) is an AI system and its output. Epistemically speaking, we are in the presence of *two*

---

[265] Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents," *cit. supra*, at 125.

[266] See *supra*, Sub-Section 2.1.

sets of accuracy standards: those that apply to AI output and those that apply to the explanations pertaining to that output.

In AI scholarship, accuracy has been closely associated with *performance*. According to Liang *et al.*, it "highlights any performance benefits of relying on the recommendation and offers a benchmark against which individuals can judge their own performance."[267] Alternatively, explanations pertaining to AI output "are able to measure the importance of parts of the input or intermediate features towards a model's decision - and can therefore be viewed as an additional and high-dimensional measurement for the discussed properties, depending on the application."[268] In AI jargon, explanations are meant to allow "for a better (compared to, e.g., just relying on the prediction error) control of the model behavior."[269]

The oft-recalled trouble with advanced ML systems is *opacity*. As Edwards and Veale put it, AI technologies may exhibit implicit rather than explicit logics since the ways in which they learn about, and shape reality do not often offer the opportunity to backtrack the stages of their inferential process.[270] Inscrutability of ML and DL models is an epistemic concern, where explanations and understanding are considered as central epistemic virtues.[271] This inscrutability is - Duede points out - that the relationship between an ML or DL model and the real world is *mediated by the logic of what the system learnt:* "no direct causal connection between the world and the DLMs mediates the model's output of a given value."[272]

To illustrate this: say a recruitment algorithm was programmed based on a simple 'if-then' rule.[273] The application of this rule would allow the system to view factors (education, work experience, career advancement, languages spoken etc) as indicators of work performance and, based on those, it would be able to infer a person's level of skill. Suppose that, when processing data not seen during training, the system - somehow - associated gender with work performance concluding that, because men's professional advancement is historically more common, they must be more skilled than women.[274] The consequent inference would be that gender is a sign of high work

---

267 Garston Liang, Jennifer F. Sloane, Christopher Donkin, Ben R. Newell, "Adapting to the algorithm: how accuracy comparisons promote the use of a decision aid" (2022) 14 *Cognitive Research: Principles & Implications,* 1, at 2.

268 Leander Weber, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement" (2023) 92 *Information Fusion*, 154, at 165.

269 *Ibid*.

270 Lilian Edwards, Michael Veale, "Slave to the Algorithm: Why a Right to an Explanation is Probably Not the Remedy You Are Looking for" (2017-2018) 18 *Duke L. & Tech'y Rev.*, 8, at 25.

271 Eamon Duede, "Instruments, agents, and artificial intelligence: novel epistemic categories of reliability" *cit. supra*, at 491.

272 *Id*., at 500.

273 If-then models are typical of so-called conditional reasoning consisting in matching a set of conditions (if a person has university education) with consequences that follow from those conditions (then the person is a highly qualified worker). Our explanation here seeks simplicity, a critical analysis of the theory of conditional reasoning being beyond the scope of this paper. For such analysis, see e.g. Ruth M.J. Byrne, Philip N. Johnson-Laird, "'If' and the problems of conditional reasoning" (2009) 7 *Trends in cognitive sciences*, 282.

274 On the effects of gendered AI systems, see Lena Wang, "The Three Harms of Gendered Technology" (2020), 24 *Australasian J. Inf. Systems*, 1.

performance *i.e*. one that the labour market favors. Amazon's gender-discriminating recruitment system provides the topical example of this.[275]

The problem with this scenario is that the gender-skill association, made by a system in its 'discovery' of the real world, may not always be foreseen by the users and even programmers.[276] This has its importance in the context of harm (unfair biases, physical injuries, illegal investments, medical misdiagnosis etc). As a matter of principle, AI-related harm is usually thought to be the result of miscalculation, error, deviation from that for which the system was trained to do.[277] The question is: how to *causally explain* the occurrence of such harm? Realist epistemic currents do not help much in answering this question. Their postulate is, essentially, that the objects of cognition are tangible occurrences with relatively discernable causes: if snow falls, we may - as some philosophers have - engage in extensive debates on the conditions under which we may assert that 'snow is white.'

In our recruitment hypothetical, the real or tangible occurrence (the AI output) does not seem to reveal a lot on the causal interrelationship (in the form of variable-association) underlying it. This leads to an important epistemic consequence. Kitchin[278] commented that, in pre-AI times, the operative assumption was that any scientific hypothesis could be tested and verified.[279] This paradigm - he argued - consisted of "overly sanitized and linear stories of how disciplines evolve, smoothing over the messy, contested and plural ways in which science unfolds in practice."[280] AI disrupted this 'sanitized' view, upsetting epistemologists' *penchant* for methodological reliability, expressed in the belief that procedures designed to produce knowledge *reliably produce* the knowledge they are designed for. In this context, is AI's capacity for knowledge-construction different from (or more sophisticated than?) that of human 'knowers'? The answer is no… and yes.

2. The Specificity (and Interpretability) of AI 'Knowledge'

The answer to the above-mentioned question ('is AI's capacity for knowledge-construction different from, or more sophisticated than, that of human 'knowers?') is

---

275 Roberto Iriondo, "Amazon scraps secret AI recruiting tool that showed bias against women," Carnegie Mellon University (available on: https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html, last visited, 20 Jan. 2024).

276 See Weston Kowert, "The Foreseeability of Human-Artificial Intelligence Interactions" (2017) 1 *Texas L. Rev.*, 181, at 204: "once the artificial intelligence is sent off to the buyer, the programmer no longer has control and the artificial intelligence could be shaped by its new owner in uncountable ways."

277 Schiyong and Kaizhong essentially view error as a judgment made in application of rules that in a given 'universe of discourse' allow to identify erroneous definitions or assertions. See Liu Shiyong, Guo Kaizhong, *Error logic: paving pathways for intelligent error identification and management* (Springer:2023), at 2-3. Chanda's and Banerjee's definition of error is more functional in the sense that they define errors in reference to the objectives (and expected outputs) of AI systems. For them, errors are 'inadequacies' which can be of two kinds: errors of comission (doing something that should not have been done) and errors of omission (not doing something that should have been done). See Sasanka Sekhar Chanda, Debarag Narayan Banerjee, "Omission and commission errors underlying AI failures" (2022) *AI & Society*, 1, at 1. In short, errors (like unfair biases) are deviations from a model's basic programming.

278 Rob Kitchin, "Big Data, new epistemologies and paradigm shifts" (2014) 1 *Big Data & Society*, 1.

279 *Id*., at 3.

280 *Ibid*.

'no' because, as already mentioned, AI is programmed based on human reasoning models, only - or so the story goes - they seem to apply those models in ways that average human agents do not.

If one lends an ear to some mainstream narratives, the attractiveness of AI stems precisely from its ability to outperform humans.[281] To a certain degree, this holds. In the field of medicine e.g., Arno *et al*.[282] sought to determine if the accuracy of AI-assisted risk-of-bias detection was comparable (noninferior) to human-only assessments. They found that in terms of *efficacy* - essentially the margin of statistical error between automated and human-only assessments - AI reached an accuracy threshold of 0.89/1 whereas for humans, the threshold was of 0.90/1.[283] AI-assisted decisions were therefore not inferior to human decisions in terms of efficacy but - the authors point out - efficacy is not an indicator of *effectiveness*, understood as the possibility for AI to produce the output that is not only accurate, but *desired* in real-life contexts. Think of the recruitment AI: if the system found that, historically, part-time workers are mostly female - which may be statistically correct - it should not be programmed to make the generalization that all women underperform in comparison to men. In this scenario, an efficacious output (though backed by statistical data) will not necessarily be viewed as effective, as it would possibly lead to restricting access to work for women, causing a text-book example of gender discrimination.

These observations allow us to fine-tune the concept of AI accuracy flagged at the beginning of this Sub-Section: although this concept is linked to the quality of AI's probabilistic reasoning, *it does matter* how this reasoning will impact the reality of humans. A well performing (accuracy-apt) system is one that would achieve a difficult double task: be statistically correct (efficacious)[284] and value-conform (effective). In this regard, regulators and scholars seem to have reasoned in terms of another *procedural parallelism*: the *design* of the inception procedures of AI systems directly shapes those system's *aptitude for accuracy*. In terms of cognition, the way knowledge about the world is *represented* in the coding phase of AI will shape the way in which AI will subsequently 'know' and 'act' in the world. In this context, it is not very surprising that regulatory and savant attention turned to the criteria used for the establishment of *ground-truths*, as a form of proto-knowledge comprised of data that an AI system can refer to when confronted with new data that is, data not seen during training.[285]

---

[281] See Katja Grace, Allan Dafoe, Baobao Zhang, Owaian Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts" (2018) 62 *J. of AI Res.*, 729.

[282] Anneliese Arno, James Thomas, Byron Wallace, Iain Marshall, Joanne E. McKenzie, Julian H. Elliot, "Accuracy and Efficiency of Machine Learning-Assisted Risk-of-Bias Assessments in 'Real World' Systemic Reviews: A Noninferiority Randomized Controlled Trial" (2022) 7 *Annals of Internal Medicine*, 1001.

[283] *Id*., at 1004.

[284] Efficacy is essentially a matter of accurate representation, not only of concrete outputs, but also of how accurately AI systems represent their targets. See Eamon Duede, "Instruments, agents, and artificial intelligence : novel epistemic categories of reliability" *cit. supra*, at 496.

[285] Lebovitz *et al.* define the term 'ground truth' as referring to the labels assigned to the data sets used to train a ML model to link new inputs to outputs and to validate its performance. See Sarah Lebovitz, Natalia Levina, Hila Lifshitz-Assa, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What" (2021) 3 *MIS Quart'y*, 1501, at 1509.

It goes without saying that the selection of the data used to constitute ground truths should be performed with great caution in the protocolized process called *labelling*: the assembling and 'cleaning' of data used during a model's programming.[286] Ground data constitutes the *cognitive referent* the system will use when performing in practice. To assess the quality of this performance, the model undergoes *training* that is, the phase where it is confronted to sub-sets of preselected data. If the model performs well (*i.e*. the risk of error is minimal or 'tolerable'), the model would go on to the so-called *validation stage*.

It should be mentioned that a well performing AI is never bias-free, but one that arrives at statistically accurate outcomes *in spite of the biases* that may be either embedded in the ground data or learnt during the model's lifetime. We have discussed elsewhere that accuracy, in AI jargon, is really a balance between *bias* (preferences embedded in the ground data) and *variance* (a model's ability to make relevant decisions and predictions when confronted to data not seen during training).[287] This balance is struck through much testing and controlling of the sample size used in the training stage. With accuracy as *fil rouge* of this paper, we will rather focus on the epistemic conditions that usually warrant 'accurate' AI output. In this vein, ground truths play the role of *premises* the accuracy of which should, logically, dictate the accuracy of the conclusions.

This is the underlying *leitmotiv* of labelling: once ground truths are selected, the systems are trained to create associations between variables, generating a series of relative weights that can be applied to future data inputs.[288] Lebovitz *et al.* refer to - what they view as - a standard method of measuring the quality of an AI model which involves the calculation of how often the model's predicted outputs match the label *a priori* defined as accurate in the data set reserved for model validation.[289] This assessment of course requires expertise, but not only. The authors cite radiology as an example: professionals in this field are trained to refer to the 'Area Under Curve' (AUC) when determining if any technological tool (ranging from imaging equipment to analytical tools) improves diagnostic accuracy.[290] AUC is therefore "primary evidence of performance"[291] supported by larger scientific acceptance (expertise published in specialized journals e.g.) and combined with other methods available for the accuracy

---

[286] Carbonara and Sleeman sfocus on the process of knowledge construction for the purpose of AI programming. For any knowledge-based system - they argue - the process of *accurate representation* of domain knowledge includes three main stags: *knowledge elicitation*, *knowledge representation* and *testing/refining* the initial knowledge base (KB_0). In the first two stages consist in using various automated tools for knowledge elicitation and representation. Knowledge refinement is a process through which the initial knowledge base KB_0 is tested and fine-tuned. To do so, two sets of cases are used: training cases used for knowledge refinements and training cases used to measure the effectiveness of those refinements, thus allowing to measure a system's effectiveness and performance in practice. See Leonardo Carbonara, Derek Sleeman, "Effective and Efficient Knowledge Base Refinement" (1999) 37 *ML*, 143, at 144.

[287] Ljupcho Grozdanovski, "In Search for Effectiveness and Fairness in Proving Algorithmic Discrimination in EU law" (2021) 58 *CMLREv.*, 99, at 107.

[288] Sarah Lebovitz, Natalia Levina, Hila Lifshitz-Assa, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *cit. supra*, at 1503.

[289] *Ibid*. The calculation is represented by a metric called the 'Area Under the Receiver Operating Curve' (AUC) and plotted on two-dimensional graphs. The AUC is a summary of a model's success and error rates, with predictions of possible false negatives and false positives. See *ibid*.

[290] *Id*. at 1508.

[291] *Ibid*.

assessment of a given system. This suggests that AI programming is integrated in broader scientific and social contexts with already existing methods of seeking and verifying information: "in knowledge-intensive contexts, experts developed over the years rich know-how practices to form high-quality knowledge outputs."[292]

Because expert fields and new technologies evolve side by side, coding should be an extremely cautious process when it comes to 1. deciding which data is 'true enough' (at a given point in time) to be used as ground data; 2. embedding models of reasoning that can allow a system to rely on that data and produce an accurate (*i.e.* efficacious *and* effective) outcome.[293] Of course, high-quality, bias-free ground data gives some assurance that a system will perform well when 'released into the wild,' but this assurance is not absolute certainty. There is always a margin of doubt that an AI system may not produce the type of output it was programmed to produce.

This unpredictability is, arguably, why AI technologies upset standard epistemology (the 'yes' answer to the question mentioned earlier): the absence of unfair biases in the labelled data does not automatically imply that a system's output will *systematically* be bias free.

The fact that we can no longer reliably assume the input/output parallelism (in terms of accuracy) is a sign of a much deeper epistemic shift triggered by Big Data. Indeed, the possibilities for various scientific and non-scientific communities to interact within - to borrow Floridi's jargon - the *infosphere*[294] hold the remarkable potential of increasing the speed with which (valid) knowledge is produced and disseminated. In addition, the sheer volume of Big Data presents several epistemic advantages: it can capture a whole domain and provide full resolution; there is no need for *a priori* theories, models or hypothesis for knowledge to be - as it were - distilled from the vast volumes of data; through the application of *agnostic* data analysis, the data can speak for themselves free of human bias; any patterns and relationships within Big Data are (presumed to be) meaningful and truthful; learning transcends context or domain-specific knowledge, thus can be interpreted by anyone who can code a statistic or data visualization…[295]

In this context, scholars have detected the "*troubling disconnection* between ML-based AI quality measures that were based solely on know-what aspects of knowledge and the rich know-how practices experts rely in their daily work."[296] This of course had a profound implication on the ability to assesses a system's potential risks and benefits.[297] If the process (the 'how') preceding an output could not be sufficiently explained based on output alone, quality measures needed to be put into place for in-depth assessments to be made possible. In the trials conducted by Lebovitz *et al*., the

---

292 *Id*., at 1512.

293 *Id*., 1513-1514: "to evaluate AI outputs, managers began reflecting on the know-how practices that enable internal experts to grapple with uncertainty in their daily work and produce high-quality judgments."

294 Luciano Floridi, "Ethics after the Information Revolution" *in* Luciano Floridi (ed.), *The Cambridge Handbook of Information and Computer Ethics* (CUP, 2012), 3, at 6.

295 Rob Kitchin, "Big Data, new epistemologies and paradigm shifts" 1 Big Data & Society (2014), 1, at 4.

296 Sarah Lebovitz, Natalia Levina, Hila Lifshitz-Assa, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *cit. supra*, at 1514 (emphasis added).

297 *Ibid*.

qualifications of the labelers were under high scrutiny, as was the "taken-for-granted representations of knowledge."[298] This eventually led to admitting that "even labels generated by experts limited [the] evaluations since experts' knowledge outputs were subject to deep underlying uncertainty and ignored know-how aspects of knowledge that were essential to producing knowledge in practice."[299]

In light of the above, it was a set of *professional standards* established, not so much as guaranteeing AI accuracy, but as supporting the belief - namely of users - that *accuracy was likely*.[300] In the AI Act, the accuracy-enhancing (and, trust-engineering) standards target, in particular, the so-called high-risk systems. Interestingly - but understandably - accuracy is seen as a *byproduct of resilience*. For example, Article 15 AI Act states said systems should be resilient as regards "errors, faults or inconsistencies that may occur within the system or the environment in which it operates, in particular due to their interaction with natural persons or other systems."[301] They will also be resilient with regard to attempts by unauthorized third parties to alter their use or performance by exploiting the system vulnerabilities.[302]

The technical solutions to address AI specific vulnerabilities shall include - the AI Act states - measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws.[303] In essence, high-risk AI systems should be resilient to anything that might cause them to deviate from their purpose. Whether this level of resilience can be achieved through technical standardization is an issue we have explored elsewhere.[304] At this stage, the takeaway from our observations on accuracy is that as a *concept*, as an *aptitude* (of a model) and as a *property* (of both ground data and AI output) *perfect accuracy* is technically difficult to instill and comes with no guarantees: try as they might, AI programmers are seldom in a position where they can predict that a well-performing AI system will invariably hit the mark in producing perfectly efficacious and effective output. This is a constant not only in discourse on expert systems (by now associated with the 'stone age' of AI development)

---

298 *Ibid*.

299 *Ibid*.

300 Commenting on the regulatory discourse on trustworthy AI and the use of technical standardization as the means to make AI 'trustworthy', Laux *et al*. stress the possibility that standardization is meant to 'engineer' trust. See Johann Laux, Sandra Wachter, Brent Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk, (2023) *Regulation & Governance*, 1-30, at 2.

301 AI Act*, cit supra,* Art. 15-3.

302 *Id*., Art. 15-4.

303 *Id*., Art. 15-4.

304 Ljupcho Grozdanovski, "The ontological congruency in the EU's data protection and data processing legislation: the (formally) risk-based and (actually) value/rights-oriented method of regulation in the AI Act" *cit. supra.*

but also with generative AI. Much like its more primitive predecessors, ChatGPT was also found to produce output 'tainted' by un unfair bias.[305]

In causal explanatory contexts, the million-dollar question is, of course, *why*? To give a plausible answer to this question there seem to be two sets of conditions: 1. that a given output *lends itself* to an explanation (explainability-as-interpretability); 2. that the explanation provides *adequate understanding* of the process through which that output was produced (explainability proper).

B. Accuracy Standards for Explanations of AI Output

A key doctrinal referent in this sub-section is the remarkable study produced by Barredo Arrieta *et al*.[306] on XAI where the authors highlight five operative concepts. First, *understandability* or *intelligibility*, which denotes "the characteristic of a model to make a human understand its function - how the model works - without any need for explaining its internal structure or the algorithmic means by which the model processes data internally."[307] Second, *comprehensibility* which "refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion."[308] Third, *interpretability* defined as "the ability to explain or to provide the meaning in understandable terms to a human."[309] Fourth, *explainability*, "association with the notion of explanation as an interface between humans and a decision maker (and is) at the same time, both an accurate proxy of the decision maker and comprehensible to humans."[310] Finally, *transparency*: "a model is considered transparent if by itself it is understandable."[311]

We can derive from the relevant scholarship that, in the field of AI, explainability can be *a priori* or *ex post*. *A priori* (*ad hoc*) explainability pertains to the criteria or standards which, if followed, are assumed to, if not guarantee, at least contribute to a system's explain-*ability* down the line (**Sub-Section 3.2.1.**) *Ex post* (*post hoc*) explainability pertains to the interpretation (retro-rationalization) of AI output, once such output is produced (**Sub-Section 3.2.2.**).

---

305 A recent study analyzing the output of two large language models (LLMs) namely ChatGPT and Alpaca, charged with drafting recommendation letters for hypothetical workers. It was observed that the language used by both systems to describe the workers was heavily gendered (using 'expert' and 'integrity' for men and 'beauty' or 'delight' for women). See Christ Stokel-Walker, "ChatGPR Replicates Gender Bias in Recommendation Letters" available on: https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/#:~:text=But%20a%20new%20study%20advises,recommendation%20letters%20for%20hypothetical%20employees (last accessed on 20 Jan. 2024).

306 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (2020) 58 *Information Fusion*, 82.

307 *Id*., at 84.

308 *Ibid*.

309 *Id*., at 85.

310 *Ibid*.

311 *Ibid*. In their study, Barredo Arrieta et al. divide transparent models into three categories: simulatable, decomposable and algorithmically transparent.

1. Ad Hoc Explainability: Embedding Transparency, Hoping for Explicability

The object of *ad hoc* explainability is a matter of standardization, essentially translating to the observance of pre-established functional and operational requirements meant to enhance a model's comprehensibility.[312] This is, no doubt, the reason why technical standardization was ultimately favored by the EU's legislature in regulating AI systems. The 'standardization narrative 'can be traced back to the HLEG's Ethics Guidelines[313] where explicability appears as one of the four cardinal principles for ethical AI, alongside the respect for human autonomy, prevention of harm and fairness. This principle - the experts argued - is crucial for building and maintaining trust in AI.[314] Curiously, the HLEG distinguished between *explicability* and *explainability*.

According to the Guidelines, *explicability* refers to the factors that support and reinforce it. Those factors are unsurprising: transparency and clarity of communication.[315] Where explicability is obstructed, the HLEG stressed that other measures (e.g. traceability, auditability and transparent communication on system capabilities) can be required, "provided that the system as a whole respects fundamental rights."[316] Alternatively, *explainability* is a *component of transparency*, pertaining to the "ability to explain both technical processes of an AI system and the related human decisions." [317] In connection to explainability, the HLEG emphasized human understandability[318] derived from explanations of the degree to which an AI system influences and shapes the decision-making process, design choices of the system and the rationale for deploying it.[319]

The distinction between explicability and explainability in the HLEG's Guidelines is interesting. Explicability seems to refer to the factors (transparency and clarity) that support a model's *interpretability*. From the vantage point of explanatory epistemology examined previously, it is possible to argue that those factors are meant to support an explanation's *objectivist* dimension or facticity.[320] In other words, transparency and clarity should make - what in a legal setting would be considered as - *elements of fact* (ground data, programming, training and validation etc) discernable, so that a model's functioning and output can *in fine* be interpreted. Alternatively, explainability - as the HLEG seems to understand it - is more *subjectivist*, explainee-oriented, focused on the *format* and *features* that explanations must have to be *understandable*.

---

312 According to Guidotti *et al*., the functional requirements of XAI are those that identify the algorithmic adequacy of a particular approach for a specific application, while operational requirements take into consideration how users interact with an explainable system and what is the expectation. See Ricardo Guidotti, Anna Monreale, Dino Pedreschi, Fosca Giannotti, "Principles of Explainable Artificial Intelligence" in Moamar Sayed-Mouchaweh (ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications* (Springer, 2021), 9, at 12.

313 High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI* (2019), *cit. supra*.

314 *Id*., at 13.

315 *Ibid*.

316 *Ibid*.

317 *Id*., 18.

318 *Ibid*.

319 *Ibid*.

320 See *supra*, Sub-Section 2.1.1.

We consider that the explicability/explainability distinction in the HLEG's Guidelines is an issue of semantics. As will be argued further, XAI is multilayered. However, concepts such as interpretability, comprehensibility and transparency are *instrumental* to explainability as the generic, operative concept in the field of XAI. In light of this, in the remainder of this paper, we will not use the HLEG's explicability/explainability distinction but will instead generically use explainability in our analysis of both the factive and subjective aspects of explanations pertaining to AI performance and outpout. Semantic parenthesis closed.

Following the HLEG's Guidelines, the AI Act translated the requirements on explainability in technical standards targeting, in particular, the so-called high-risk systems. These can be clustered in roughly *three families*.

The *first* includes standards that generate *requirements for accuracy* (of the ground data) and *transparency*. These requirements pertain to data governance and management practices such as relevant design choices,[321] data collection;[322] relevant data reparation processing operations, such as annotations, labeling, cleaning, enrichment and aggregation,[323] the formulation of relevant assumptions, namely with respect to information that the data are supposed to measure and represent,[324] prior assessment of the availability, quantity and suitability of the data sets that are needed,[325] examination in view of possible biases [326] and identification of data gaps or shortcomings, and how those can be addressed. [327] Unsurprisingly, the AI Act expresses a basic requirement that training, validation and testing data sets be *relevant*, *representative*, *free of errors* and *complete* [328] taking into account, "to the extent required by the intended purpose" the characteristics pertaining to specific geographical, behavioral and functional setting within which the high-risk system is intended to be used. [329] The *data governance requirement* is, of course, meant to increase the transparency and provision of information to users. Article 13(1) states that high-risk AI systems shall be "designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately." Perhaps naively, this Article states that the "appropriate type and degree of transparency" - whatever 'appropriate' is - will be reached through compliance with the obligations set out in the AI Act.[330] High-risk systems shall, in addition, be accompanied by instructions for use, in a digital format, that include concise, complete, correct and clear information that is "relevant, accessible and comprehensible for users." [331] The information required includes *inter alia* the characteristics, capabilities and limitations of performance of the high-risk system including its intended purpose,[332] the levels of accuracy robustness and cybersecurity against which the system had been tested and validated and which "can be expected"

321 AI Act, *cit. supra*, Art. 10-2 (a).
322 *Id*., Art. 10-2 (b).
323 *Id*., Art. 10-2 (c).
324 *Id*., Art. 10-2 (d).
325 *Id*., Art. 10-2 (e).
326 *Id*., Art. 10-2 (f).
327 *Id*., Art. 10-2 (g).
328 *Id*., Art. 10-3.
329 *Id*., Art. 10-4.
330 *Id*., Art. 13-1.
331 *Id*., Art. 13-2.
332 *Id*., Art. 13-3(b)(i).

as well as any known and foreseeable circumstances that may have an impact on the expected level of accuracy, robustness and cybersecurity,[333] any known or foreseeable circumstance related to the use of a high-risk system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety of fundamental rights,[334] its performance as regards the persons or groups on which the system is intended to be used,[335] when appropriate, specifications for the input data, or any relevant information in terms of training, validation and testing data sets used, taking into account the intended purpose of the AI system.[336] The information should further include the *changes* of the high-risk AI system determined by the provider during the initial conformity assessment,[337] the human oversight, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users,[338] the expected lifetime of the high-risk system and any necessary maintenance and care measures to ensure the proper functioning of that system, including as regards software updates.[339]

The *second family* of standards create requirements to *produce proof of compliance* and *traceability*. Under these requirements, the programmer is held to keep technical documentation,[340] drawn up "in such a way to demonstrate" the compliance of a high-risk AI system with the AI Act. They should also perform record-keeping able to show that high-risk systems are designed with capabilities enabling the automatic recording of events ('logs') while those systems are operating.[341] The logging capabilities should increase the level of traceability[342] and facilitate monitoring of a system's operation in situations where it may present a risk of harm.[343] In a similar vein, Article 11(4) of the AI Act states that the logging capabilities should provide, "at a minimum" recording of the period of each use of a given system,[344] the reference database against which input data has been checked by the system,[345] the input data for which the search has led to a match[346] and the identification of natural persons involved in the verification of the output.[347]

The *third family* of standards pertain to *human oversight*. Article 14 of the AI Act creates the obligation to provide appropriate human-machine interface tools so that high-risk AI systems can be effectively overseen by natural persons during those systems' use.[348] It should prevent and minimize the risks to health, safety or fundamental rights that may emerge during the intended use of the AI system or in conditions of reasonably foreseeable misuse.[349] In a positive sense, human oversight

333 *Id*., Art. 13-3(b)(ii).
334 *Id*., Art. 13-3(b)(iii).
335 *Id*., Art. 13-3(b)(iv).
336 *Id*., Art. 13-3(b)(v).
337 *Id*., Art. 13-3(c).
338 *Id*., Art. 13-3(d).
339 *Id*., Art. 13-3(e).
340 *Id*., Art. 11.
341 *Id*., Art. 12-1.
342 *Id*., Art. 12-2.
343 *Id*., Art. 12-3.
344 *Id*., Art. 11-4(a).
345 *Id*., Art. 11-4(b).
346 *Id*., Art. 11-4(c).
347 *Id*., Art. 11-4(d).
348 *Id*., Art. 14(1).
349 *Id*., Art. 14(2).

should be ensured through measures such as identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service,[350] identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the user.[351] These measures are meant to enable individuals to whom human oversight is assigned to fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;[352] remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;[353] be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;[354] be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;[355] be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure.[356]

No doubt for convenience, the rationale which transpires from these 'families of standards' is one of *epistemic parallelism* by virtue of which *procedures designed to increase AI accuracy should yield accurate and explainable outcomes*. But is this parallelism tenable? Though several factors can explain the EU's *penchant* for standardization, it is open to criticism on namely *three points*: 1. the technical standards are descriptive and vaguely worded. Presumably, even if the AI Act did not set out a duty of transparency, software engineers would still abide by it as a *deontic requirement* in their sector of activity; 2. the procedures/outcomes parallelism as underlying rationale of the AI Act is somewhat naïve. Bearing in mind our observations on the epistemology of AI knowledge construction,[357] there are no absolute guarantees that systems' conformity to technical standards will prevent them from 'deviating' from their original programming; 3. the parallelism assumption seems to have shaped regulators' view of how to achieve explaninability. The propositional (if/then) logic that characterizes this view can be summarized as follows: *if* there is compliance with the standards in the AI Act *then* AI output is accurate and explainable (statement labelled as true); a natural or legal person has complied with the AI Act (premise), a system's output is surely accurate and explainable (conclusion).

The peculiarity of this reasoning is that explainability becomes a *byproduct of lawfulness*. On the one hand, this is not surprising. When legislation includes series of technical standards, those are presumably drawn from existing business practices of, say, manufacturing a specific type of products. Through their translation into law, those standards acquire the authority of the law and generate mandatory requirements which serve as referents for the assessment of the legality of market actors' conduct.

---

350 *Id*., Art. 14-3(a).
351 *Id*., Art. 14-3(b).
352 *Id*., Art. 14-4(a).
353 *Id*., Art. 14-4(b).
354 *Id*., Art. 14-4(c).
355 *Id*., Art. 14-4(d).
356 *Id*., Art. 14-4(e).
357 See *supra*, Section 2.

On the other hand however, the argument of lawfulness is not fully satisfying for the purpose of giving fact-of-the matter causal explanations. What victims *need*, in terms of understanding, is an explanation of how a system operating in a specific context developed, say, a bias. This bias may, of course, be the consequence of non-compliance with the AI Act, but it may occur even when the standards in this instrument were religiously observed. Selecting lawfulness as the be-all-end-all factor for accurate AI output is too limiting in cases where the cause of AI-related harm may reside with a system having acted alone. *Ad hoc* explainability provides understanding on what *ought to be done* for AI output to be explainable; it does not necessarily deliver understanding on the decisional process that led to an output which failed to be explainable. For that type of understanding to be given, *post hoc* explainability is paramount, translating to several (some sophisticated and complex) explanatory methods and techniques experts apply once - possibly harmful - AI output has been produced.

2. Post-Hoc Explainability: Experiencing Opacity, Attempting Explanation

The impression one has when reading the AI Act is that of a binary view of explainability: a system is either created transparent and is therefore explainable, or it is not. In software engineering, explainability, especially *post hoc* explainability is a *spectrum*. The nature and feasibility of *post hoc* explanations are largely dictated by the complexity of the models used in the programing of AI systems. The general rule of thumb is not difficult to understand: the more 'linear' the model (*i.e.* where the association between variables is continuous), the more transparent and explainable the system. From the perspective of AI programming, there are several techniques available:

text explanations,[358] visualizations,[359] local explanations,[360] explanations by example,[361] explanations by simplification[362] and feature relevance.[363]

Barredo Arrieta *et al*.[364] produced a well-documented study showcasing the various reasoning models and corresponding levels of explainability. There are, indeed, models that can reliably be qualified as transparent are explainable. They generally apply linear/logistic regression[365] meaning that they are rule-based and operate on the assumption of a linear dependence between predictors and predicted variables. They are 'stiff' as they do not tend to deviate from the rules which makes them predictable and transparent and their output *prima facie* explainable. This family of explainable models includes *inter alia decision trees* which are hierarchical structures used to support regression and classification. Guidotti *et al*.[366] explain that decision trees exploit a graph-structure with so-called internal nodes representing tests on features or attributes (e.g., whether a variable has a value lower than, equal to, or greater than a threshold) and so-called leaf nodes representing a decision. Each 'branch' is a possible outcome. The connections from the 'root' to the 'leaves' represent the so-called

---

358 Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Expainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 88: "text explanations deal with the problem of bringing explainability for a model by means of learning to generate *text explanations* that help explaining the results from the model. *Text explanations* also include every method generating symbols that represent the functioning of the model. These symbols may portrait the rationale of the algorithm by means of a semantic mapping from model to symbols."

359 *Ibid*: "Visual explanation techniques for post-hoc explainability aim at visualizing the model's behavior. Many of the visualization methods existing in the literature come along with dimensionality reduction techniques that allow for a human interpretable simple visualization. Visualizations may be coupled with other techniques to im- prove their understanding, and are considered as the most suitable way to introduce complex interactions within the variables involved in the model to users not acquainted to ML modeling."

360 *Ibid*: "local explanations tackle explainability by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model. These explanations can be formed by means of techniques with the differentiating property that these only explain part of the whole system's functioning."

361 *Ibid*: "Explanations by example consider the extraction of data examples that relate to the result generated by a certain model, enabling to get a better understanding of the model itself. Similarly to how humans behave when attempting to explain a given process, *explanations by example* are mainly centered in extracting representative examples that grasp the inner relationships and correlations found by the model being analyzed."

362 *Ibid*: "*Explanations by simplification* collectively denote those techniques in which a whole new system is rebuilt based on the trained model to be explained. This new, simplified model usually attempts at optimizing its resemblance to its antecedent functioning, while reducing its complexity, and keeping a similar performance score. An interesting byproduct of this family of post-hoc techniques is that the simplified model is, in general, easier to be implemented due to its reduced complexity with respect to the model it represents."

363 *Ibid*: "feature relevance explanation methods for post-hoc explainability clarify the inner functioning of a model by computing a relevance score for its managed variables. These scores quantify the affection (sensitivity) a feature has upon the output of the model. A comparison of the scores among different variables unveils the importance granted by the model to each of such variables when producing its output. *Feature relevance* methods can be thought to be an indirect method to explain a model."

364 *Id*., at 82.

365 *Id*., at 88-90.

366 Ricardo Guidotti, Anna Monreale, Dino Pedreschi, Fosca Giannotti, "Principles of Explainable Artificial Intelligence" in Moamar Sayed-Mouchaweh (ed.), *Explainable AI Withiin the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications* (Springer, 2021) 9.

classification rules. The most common rules are the conditional if-then rules, where the 'if' clause provides a set of conditions on the input variables. If the conditions are met, the system proceeds to drawing a corresponding conclusion (the 'then' portion of the reasoning). For a list of rules, the AI "returns as the decision the consequent of the first rule that is verified. Linear models allow visualizing the *feature importance*: both the sign and the magnitude of the contribution of the attributes for a given prediction."[367] In the simplest of their flavors - Barredo Arrieta *et al.* write - trees are simulatable models, manageable by human agents: "many applications of these models fall out of the fields of computation and AI (…) meaning that experts from other fields usually feel comfortable interpreting the outputs of these models."[368] However, the authors stress that decision tress have poor generalization properties which make them less interesting for businesses. Instead, so-called *K-Nearest Neighbors (KNN)* are more attractive.

KNN learning "combines the target values of K selected neighbors to predict the target value of a given test pattern."[369] When predicting a class of a test sample, they refer to classes of its K nearest neighbors (the 'neighborhood' relation being function of distance between samples).[370] KNN models work by association, much like humans who 'learn' from new experiences by associating them to similar past experiences.[371] When confronted to new sets of data, KNN models classify them in categories of the basic dataset that are similar to the data unseen during training. The simplest use of these models is e.g. that of pattern/image recognition.[372] In principle, they are predictable and explainable, which means that, to determine why a new sample has been classified inside a group, an explainer would need to refer to that sample's neighbors to infer how a 'new' sample interacted with those.[373]

In the class of linear models, Barredo Arrieta *et al.* further mention *rule-based learning*. The systems programmed with this method generate rules to characterize the data they learn from. Those rules can be linear (e.g. if-then) or combinations of such rules. So-called *fuzzy rule-based systems* enable the definition of verbally formulated rules over imprecise domains.[374] The specificity of fuzzy reasoning models is that they depart from the standard true/false dichotomy. Propositional logic typically offers a binary view: if a premise 'A' is true, the consequent 'B' is also true. Fuzzy logic deals

---

367 *Id*., at 15.

368 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *cit. supra*, 91.

369 Mahmood Akbari, Peter Jules van Overloop, Abbas Afshar, "Clustered K Nearet Neighbor Algorithm for Daily Inflow Forecasting" (2011) 5 *Water resources management*, 1341, at 1343.

370 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 91.

371 *Ibid*.

372 See e.g. Si-Bao Chen, YU-Lan Xu, Chris H.Q. Ding, Bin Luo, "A Nonnegative Locally Linear KNN model for image recognition" (2018) 83 *Pattern Recognition*, 78.

373 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 91.

374 *Ibid*.

with degrees, rather than fixed values of truth and falsity. Fuzzy systems - Barreda Arrieta *et al.* argue - empower more understandable models, since they operate in linguistic terms and perform better than classic rule systems in context with degrees of uncertainty.[375] Those systems are used e.g. in trading, in cases where traders seek to optimize portfolios while taking into consideration several factors.[376] In principle, fuzzy models are interpretable, though problems may arise when the rules they generate are too long.[377] A design goal usually sought by a user is to be able to analyze and understand the model; the number of rules in a model clearly improves its performance but also compromises its interpretability. In addition to the number of rules, their specificity may also adversely affect interpretability: a high number of antecedents and/or consequences might become difficult to interpret.[378]

In a similar vein, *Generalized additive models (GAM)* should be mentioned. They include two variables: a *response variable* (the consequent) and *predictor variables* (antecedents). They are 'linear' because their responses depend on so-called *unknown smooth functions of predictor variables*. 'Smoothness' is function of continuous derivatives in a given set called the *differentiability class*. In essence, continuous derivates are sign of stability of the variables and tend to 'stabilize' the response variable. GAMs are thus able to infer the smooth functions whose aggregate composition approximates the predicted variable.[379] In principle, GAMs too are interpretable, allowing users to verify the importance of each variable and how it affects the predicted output. The last model Barredo Arrieta *et al.* cite as interpretable are *Bayesian networks*. They make links that represent the conditional dependencies between a set of variables and "fall below the ceiling of transparent models"[380] because they are simulatable, decomposable and algorithmically transparent.

Regarding the less or non-interpretable (because non-linear) models, Barreda *et al*. cite essentially *three families of models*. First, the so-called *tree ensembles*, *forests* and *multiple classifier systems*. These are - arguably - among the most accurate (in terms of efficacy) because they are assumed to improve *generalization* capability of single-decision trees which are usually prone to so-called overfitting.[381] To avoid overfitting, tree ensembles combine different trees to obtain an aggregated

---

375 *Ibid*

376 See e.g. Yong Zhang, Weilong Liu, Xingyu Yang, "An automatic trading system for fuzzy portfolio optimization problem with sell orders" (2022) 187 *Expert Systems with applications*, 115822.

377 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 91.

378 *Ibid*.

379 *Ibid*.

380 *Id.*, at 92.

381 *Id*., at 94 Overfitting refers to a case where a system's variance (essentially the ability to 'learn' from new data) is high, running the risk of the system taking into account elements that are irrelevant for the performance of a given task. Overfitting is usually thought to be the consequence of a system's exposure to noise *i.e.* irrelevant data. We have previously discussed overfitting in recruitment scenarios. A recruitment AI might 'overfit' if e.g. it considered that job applicants who do not update their LinkedIn status regularly are introverts, not fit to work in teams. This might cause the system to exclude such applicants from the recruitment process. The 'overfitting' would essentially stem from the fact the system would not prioritize hard skills to shortlist job applicants, but due to its exposure to 'noise' would consider as determining factors that might have little or nothing to do with a set of job requirements. See Ljupcho Grozdanovski, "In search for effectiveness and fairness in proving algorithmic discrimination in EU law," *cit. supra*, at 108.

prediction/regression.[382] Though overfitting can be avoided, the combination of models makes the interpretation of an overall ensemble more complex than that of each of its compounding elements, forcing the user to employ *post hoc* interpretation techniques such as simplification, feature relevance estimators, text explanations, local explanations and model visualizations. Simplification consists in the creation of a less complex model from a set of random samples from the labeled data. It can also include a so-called *Simplified Tree Ensemble Learner (STEL)* which - again - consists in using two models, one simple and one complex, the former being used to interpret the latter through so-called Expectation-Maximization and Kullback-Leibler divergence.[383]

Another technique is *feature relevance*, especially used in tree ensembles. Feature relevance consists in measuring the so-called *Mean Decrease Accuracy (MDA)* of a forest, when a certain_variable is randomly permuted in the out-of-bag samples. This method allows experts to determine how the usage of variable importance reflects the underlying relationships in a Random Forest. Finally, a so-called *crosswise technique* proposes a framework that poses recommendations which convert an example from one class to another. The idea here is to disentangle the variables' importance in a way that is further descriptive.[384]

The second type of less/non-interpretable models cited by Barreda *et al.* are the so-called *Support Vector Machines (SVM)* which are more complex and opaque than tree ensembles.[385] SVMs construct so-called hyper-planes (or a set of hyper-planes) in a high (or infinite) dimensional space, which can be used for classification, regression or other tasks.[386] The accuracy of SVM is a function of the distance (functional margin) between the hyperplane and the nearest training-data point of any class. The larger the margin, the lower the generalization error of the classifier[387] (namely because distance reduces noise and allows the classifier to 'zoom in' on relevant training data points). The techniques used to explain SVMs are simplification, local explanations, visualizations and explanations by example. Simplifications here include four classes. First, building of rule-based models from the support vectors of a training model. This approach consists in extracting rules from the support vectors of a trained SVM using a modified sequential covering algorithm.[388] This may yield fuzzy rules in lieu of standard, propositional rules.[389] The argument voiced by experts is that long antecedents reduce comprehensibility, and a fuzzy approach allows for a more linguistically understandable result.[390]

The second approach consists in adding an SVM's hyperplane, along with support vectors, to the components in charge with creating the rules. This translates to

---

382 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 94.

383 *Id*., at 94. The Kullback-Leibler divergence allows to measure the degree of dissimilarity between two probability distributions.

384 *Ibid*.

385 *Id*., at 95.

386 *Id*., at 95.

387 *Ibid*.

388 *Ibid*.

389 *Ibid*.

390 *Ibid*.

creating hyper-rectangles from the intersections between the support vectors and the hyper-plane.[391]

The third approach consists in adding the actual training data as a component for building the rules - this would translate to creating a clustering method to group prototype vectors for each class. This combination allows for the defining of ellipsoids and hyper-rectangles in the input space.[392]

The fourth method is using SVC to give an interpretation to SVM decisions in terms of linear rules that define the space in Voronoi sections from extracted prototypes.[393]

Finally, there are the *Deep Learning models* - multi-layer networks capable of inferring complex relations among variables.[394] Because of this, they are assumed to be highly performing, but also raise serious interpretability/explainability issues. The techniques used to increase explainability are model simplification, feature relevance estimators, text explanations, local explanations and model visualizations. Barredo Arrieta *et al.* cite, as an example, the Deep RED algorithm, which extends the decompositional approach to rule extraction (essentially splitting the neuron level) for multi-layer neural network by adding more decision trees and rules.

Among generally used simplification techniques, a method called *Interpretable Mimic Learning* is used to extract an interpretable model by means of gradient boosting trees. Experts propose a hierarchical partitioning of the feature space that reveals the rejection of unlikely class labels, until association is predicted.[395] Since simplification of multi-layer neural networks is increasingly complex as the number of layers increases, feature relevance methods have become more commonly used for increasing explainability. One approach here would be to decompose the network classification decision into contributions of its input elements. This would translate to considering each neuron as an object that can be decomposed and expanded then aggregate and back-propagate these decompositions through the network, resulting in a deep Taylor decomposition.[396]

The main takeaway from our brief - though technical - overview of *post-hoc* explainability is its *complexity*. Engineers seem to have quite the 'toolbox' of techniques and methods that can easily adapt to the type of model that requires explanation. However, none of the *post hoc* explainability techniques and methods magically delivers *accurate* explanations. Explanation methods as a *post-hoc* on black-box models are not 100% faithful to the original and often do not provide enough detail to understand how the black-box models are predicting.[397] Yet, *post hoc* explanations are perhaps those capable of providing the most convincing (plausible and probative) understanding of causation in AI liability cases. In other words, XAI is - or should be - a prerequisite to the litigants' ability to give to causal explanations when debating the

391 *Ibid*.
392 *Ibid*.
393 *Ibid.*
394 *Ibid*.
395 *Id*., at 96.
396 *Ibid*.
397 Uday Kamath, John Liu, E*xplainable Artificial Intelligence: An Introduction to Interpretable Machine Learning* (Springer, 2021) at 122.

origin of AI-related harm. As intuitively obvious as this might seem, legal views are diverging. The following Sub-Section will showcase that divergence by outlining three legal perspectives.

III. XAI, INTEGRAL TO CAUSAL EXPLANATIONS? THREE PERSPECTIVES

With our discussion of explanatory accuracy and accuracy in connection to XAI in the backdrop, the *relevant procedural question* is whether plausibly accurate (or believable) causal explanations require understanding provided through the explainability methods (*ad hoc* and *post hoc*) mentioned above. Intuitively, the answer would be 'yes.' After all, when harm is *occasioned* by the use of an AI system, it is only natural to seek to uncover the role the system played in that harm materializing. This suggests that the law - including EU law - should include a set of *procedural abilities* that would allow litigants to engage in a discovery of facts that would reveal: 1. the actual (as opposed to the presumed) causal power of the AI system to be established and explained; 2. the nature and the extent of the human involvement in the system's harmful output; 3. the agent who should be held to compensate the harm occasioned by the system. In sum, the law should give an appropriate response to the *epistemic needs* of litigants in AI liability cases, in order to support their meaningful (and effective) participation in the resolution of AI liability cases. But what exactly are those needs? To use explanatory jargon, *what type(s) of understanding* do litigants flag as necessary to play an active role in the adjudication process? The emerging caselaw, as well as the EU's regulation on data processing and AI liability reveals three perspectives.

In several studies of the General Data Protection Regulation (GDPR)[398] scholars have interpreted the so-called *right to a human explanation* as needing to yield understanding of the functionalities of an AI system, therefore include *post-hoc* explainability (**Sub-Section 4.1.**). Emerging North-American caselaw in AI liability gives an additional hint: the litigants in many judicial instances do indeed seek to understand how a given system worked, but they also flagged as necessary the understanding of the *reasons why reliance on a given AI output was justified* (**Sub-Section 4.2.**). Finally, there is the EU perspective which is peculiar: the understanding the forthcoming AI liability regulation will support is neither on a system's functionalities, nor on the reasons underlying the decision to rely on that system's output. The understanding said regulation will enable pertains to the level of compliance of defendants (programmers or users) with applicable technical standards such as those enshrined in the AI Act (**Sub-Section 4.3.**).

A. 'It's about Understanding How (A System Works)' - Experts Said

The GDPR does not explicitly mention a *right* to (human) explanation. It does, however, include a *provision on transparency*, as a necessary legal (and epistemic) precondition for explainability. The normative blueprint for the principle of transparency comes from Article 12 GDPR which states that "any communication"

---

398 Regulation n° 2016/679 of the European Parliament and of the Council, of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46 (General Data Protection Regulation - GDPR), *OJ n° L 119, 4.5.2016, p. 1.*

relating to the data subject should be given by the data controller in a "concise, transparent, intelligible and easily accessible form, using clear and plain language."[399] The meaning of transparency we can derive from this Article is not difficult to grasp: for data processing to be transparent, the data subject should have access to *relevant* information - whatever those are - which should be conveyed to them clearly. The Article 29 Working Party (A29WP) - the predecessor to the European Data Protection Board (EDPB) - made the additional connection between transparency, fairness and accountability. It stressed that "the controller *must always be able to demonstrate* that personal data are processed in a transparent manner in relation to the data subject."[400]

If we read the A29WP guidelines through the *lens of evidence*, the Working Party seems to place, on the controller, the *onus* of proving transparency. They should be able to meet this 'burden' in three key stages of a data processing cycle: *before* this process is launched (when the personal data is collected either from the data subject or otherwise obtained), *throughout* the data processing (when communicating with data subjects about their rights) and *at specific points* while processing is ongoing (say, when data breaches occur or in the case of material changes to the processing).[401] To 'demonstrate' transparency, data controllers are required to present information/communication "efficiently and succinctly"[402] and the information "should be clearly differentiated from other non-privacy related information such as contractual provisions or general terms of use."[403]

It should of course be mentioned that transparency in the context of the GDPR applies in the processing of *personal data* only. There is room for debate on whether 'transparency' as enshrined in said instrument is equivalent to transparency as interpreted in connection to AI (which could process both personal and non-personal data). This is a debate deserving of a separate study. For the purpose of this paper, we shall assume that Article 12 GDPR (as interpreted by the A29WP) gives the canon on how a *generic* duty of transparency should support explainability in any data processing context. Based on this assumption, let us zoom in on the application of this 'generic understanding' of transparency in the context of *automated* data processing. Article 22 GDPR is relevant here.

By virtue of said article, the data subject has the right *not to be subject* to a decision "based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[404] Exceptionally, automated data processing can be allowed in three cases: 1. for the entering into, or performance of, a contract between the data subject and the data controller;[405] 2. when such processing is authorized by Union or Member State law to which the controller is subject;[406] 3. when the decision is based on the data subject's

399 *Id*., Art. 12(1).
400 Article 29 Working Party, *Guidelines on transparency under Regulation 2016/679* (29 November 2017, last revised on 11 April 2018), available on: https://ec.europa.eu/newsroom/article29/items (last accessed on 20 Jan. 2023), at 5.
401 *Id*., at 6.
402 *Id*., at 7.
403 *Id*., at 7.
404 GDPR, *cit. supra*, Art. 22(1).
405 *Id*., Art. 22(2)(a).
406 *Id*., Art. 22(2)(b).

explicit consent.[407] In these 'exceptional' cases, the data controller is required to implement "suitable measures" to safeguard the data subject's rights, freedoms and legitimate interests, "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."[408]

Article 22 GDPR has been interpreted as integrating human explanation in an *entitlement* (right), though this provision does not at all address the content and scope of that explanation. It does however highlight its finality which is *procedural*: the explanation given should enable the data subject to 'contest the decision,' presumably in dispute-resolution procedures launched before a national data protection authority or a court. The A29WP's Guidelines on automated individual decision-making and profiling[409] shed more light on that which ought to be explained on the grounds of said Article. First, the Working Party stressed that the term 'right' (to an explanation) entails a "general prohibition for decision-making based solely on automated data processing," [410] the implication being that such processing is "not allowed generally."[411]

Second - and more interestingly - 'automated decision' according to A29WP is one that implies *no human involvement*: "to qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than a token gesture."[412] The Guidelines further state that this "should be carried out by someone who has the authority and competence to change the decision."[413] This type of decision should, moreover produce effects that "must be sufficiently great or important to be worthy of attention."[414] Typically, 'significant effects' are produced from, say, automatic refusal of an online credit application or e-recruiting practices without any human intervention. In essence the automated decision should have the potential to "significantly affect the circumstances, behavior or choices of the individuals concerned; have a prolonged or permanent impact on the data subject or at its most extreme, lead to the exclusion or discrimination of individuals."[415]

Third, the A29WP stated that the controller ought to provide *meaningful* information. To do so, they should "find *simple ways* to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision."[416] The information should however "be sufficiently comprehensive for the data subject to understand the reasons for the decision."[417] To make the explanation meaningful and understandable, "real, tangible examples of the type of possible effects should be given."[418]

---

407 *Id*., Art. 22(2)(c).
408 *Id*., Art. 22(3).
409 Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purpose of Regulation 2016/679 (3 October 2017, las revised on 6 February 2018), available on: https://ec.europa.eu/newsroom/article29/items (last accessed on 20 Jan. 2023).
410 *Id*., at 19.
411 *Id*., at 20.
412 *Id*., at 21.
413 *Ibid*.
414 *Ibid*.
415 *Id*., at 25.
416 *Id*., at 21 (emphasis added).
417 *Id*., at 25.
418 *Id*., at 25.

If a given data processing can qualify as 'automated decision' under Article 22 GDPR (as interpreted by the A29WG), there seem to be two types of requirements that stem from the right to human explanation. On the one hand, the explanations should be *holistic*, meaning that they can, or even should extend to all stages (before, during, after) of an automated decision process.[419] Wachter and Floridi[420] espoused this *holistic view*, arguing that Article 22 GDPR generated the following duties for the data processor: to give explanation *ex ante* (on how an AI system's functionalities), to give explanation *ex post* (on the rationale of a system's output) and to comply with existing legal obligations.

On the other hand, the A29WP seems to suggest the standard of *clarity* (and by that, understandability) warranted by Article 12 GDPR which mentions 'efficient and succinct' communication. The Working Party also coheres with the 'basic' epistemology of explanations by virtue of which, explanatory goodness depends on the level of understandability delivered which, of course, presupposes clarity of the explanation as such, and a satisfactory level of comprehensiveness on the side of the explainees.[421] Most importantly, and in line with the 'holistic' reading of Article 22 GDPR, the Working Group, as well as scholarship, seem to suggest that said Article should include both *ad hoc* and *pos hoc* explanations: a data subject should ideally understand a system's functionalities and the 'reasoning' pattern(s) it applied in the course of automated data processing.

B. 'It's about Understanding Why (A System is Accurate)' - Litigants Said

A shift from *understanding-how* (a system worked) to *understanding-why* (a system was relied upon) can be seen in the previously mentioned *Pickett*, *Loomis* and *Ewert*,[422] which is the Canadian *pendant* of *Loomis*. The appellant in *Ewert* challenged the use of five psychological an actuarial risk assessment tools used by the Correctional Service of Canada to assess an offender's psychopathy and risk of recidivism, on the basis that they were developed and tested on predominantly non-Indigenous populations and that no research confirmed that they were valid when applied to Indigenous persons. He claimed, therefore, that reliance on these tools in respect to Indigenous offenders breached the Corrections and Conditional Release Act. One of the issues raised in this case was that of 'reasonable steps' taken to produce accurate information about the risk of recidivism of indigenous people. The appellant argued that Canadian authorities had long been aware of concerns regarding the possibility of AI exhibiting cultural bias and yet took no action to confirm their validity, continuing to use them in respect to Indigenous offenders, despite the fact that research

---

419 See Article 29 Working Party, Guidelines on transparency under Regulation 2016/679, *cit. supra*, at 7. This 'holistic view' is also supported by Art. 68(c) (post-compromise) AI Act *cit. supra*, relative to the right to explanation of individual decision-making. Par. 1 of this provision states that "any affected person subject to a decision which is taken by the deployer on the basis of the output from an high-risk AI system listed in Annex III (…), and which produces legal effects or similarly significantly affects him or her in a way that they consider to adversely impact their health, safety and fundamental rights *shall have the right to request from the deployer clear and meaningful explanations on the role of the AI system in the decision-making procedure and the main elements of the decision taken*" (emphasis added).

420 Sandra Wachter, Luciano Floridi, Brent Daniel Mittelstadt, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation" (2017) 2 *Int'l Data Priv'y L.*, 1, at 3.

421 See our discussion on understandability *supra*, Sub-Section 2.1.2.

422 *Ewert vs. Canada*, 2018 SCC 30, File n° 37233, 13 June 2018.

would have been feasible. There is systemic discrimination against Indigenous offenders; for the correctional system to operate fairly and effectively - the appellant argued - the assumption that all offenders can be treated fairly by being treated the same way must be abandoned.[423]

The arguments in *Ewert* confirm the '*give me the reason*s' trend we also observed, namely in *Pickett*. The appellant essentially criticized the inertia of the Canadian authorities, arguing that they consistently relied on automated recidivism decisions, without even seeking to find evidence of their accuracy. We thus detect a plea for an explanation apt at delivering understanding of the *reasons* why a system should be viewed as accurate and reliable. The Canadian courts' evidentiary assessment was, however, stringent. To establish that the reliance on the automated tools violated the principle of "fundamental justice against arbitrariness" said courts argued that the appellant "had to show on a balance of probabilities that the (authorities') practice of using the impugned tools with respect to Indigenous offenders had no rational connection to the government objective."[424] The courts found he had not done so: "there was no evidence before the trial judge that how the impugned tools operate in the case of Indigenous offenders is likely to be different from how they operate in the case of non-Indigenous offenders that their use in respect of the former is completely unrelated to the government objective.' The trial judge could not have found, "on the evidence before him" that the impugned tools overestimate the risk posed by Indigenous inmates or lead to harsher conditions of incarceration or the denial of rehabilitative opportunities because of such an overestimation.[425] In other words, the appellant did not meet the standard of proof required to support his claims.

Ewert, like *Loomis*, is noteworthy. Though both cases include *requests to understand* the reasons justifying (human) reliance on AI output, they also showcase a harsh court scrutiny over the reality of the alleged harm. Whether it be gender discrimination in *Loomis*, or ethnic discrimination in *Ewert*, the courts required that the claimants present arguments (and explanations) going beyond mere suspicions or assertions. They requested that the claimants argue - ideally based on 'strong' evidence - that the systems concerned were, in fact, inaccurate. In both cases, the claimants failed to meet the standards of proof and of persuasion. Is this due to the fact that in both *Loomis* and *Ewert* a public interest (*i.e.* the functioning of national correctional systems) was at stake? Who knows. The lesson for the EU we can draw from both cases is that, in the future, defendants - which may be public or private persons - are likely to be called to: 1. give reasons for their reliance on AI output; 2. provide evidence that justify those reasons; 3. that evidence can include general expertise as well as explanations (e.g. local explanations) on a system's functionalities.

Another takeaway from the cited caselaw caselaw is that the reasons for reliance on AI output ought to be given when that output no human intervention/involvement in producing that output can be discerned. In the EU, the meaning of 'absence of human involvement' in connection to the concept of 'automated decisions' within the meaning of Article 22 GDPR, was open for debate. Finally, the *Schufa* case came along, dealing with a credit scoring system having refused the plaintiff's loan application based on the low probability that they might be able to reimburse the loan. In his Opinion, Advocate

423 *Id.*, at 169.
424 *Ibid.*
425 *Ibid.*

General (AG) Pikamäe[426] considered that the decision in this case could, indeed, qualify as automated: Article 22 GDPR does not specify *the form* that the decision should have, though its automatized nature should appear as a distinctive feature.

AG Pikamäe's position on this point is - dare we say - a reasonable one: according to him, the automated nature of a decision depends on the rules and practices of the credit establishment which should leave *no margin of appreciation* as regards the use of (and presumably, the reliance on) automated assessment tools of loan applications. In other words, automated decisions are 'automated' when they imply both *means of automated data processing* and *automatic human reliance*. The CJEU's ruling[427] however was rather laconic though generally converging with AG Pikamäe's Opinion. The Court stated that it was "common ground" that the activity of the loan-assessing private entity in *Schufa*, met the definition of profiling, as per Article 4(4) GDPR, because the automated establishing of a probability value pertaining to a person's credit related to a specific person and to that person's ability to repay a loan.[428] Interestingly, the CJEU seems to have interpreted the 'automated' portion of the 'automated decision' concept as pertaining to the *means* of personal data processing, without placing much emphasis on the 'absence of human involvement' part. In that regard, AG Pikamäe's Opinion is more elaborate.

Assuming that the AG had the right intuition on the automated human reliance aspect of automated decisions, it should be noted that the AI Act prescribes a duty of human control and oversight *prima facie* hinting to the fact that reliance should *never* be automatic. The point on which AG Pikamäe should probably have focused is the *possibility and effectiveness for ex post* human control, the relevant questions of fact being the following: 1. is a given automated decision the *determining factor* in making a final decision (e.g. approving loans)?; 2. would the human agent's decision been the same if no AI system was used? If the answer to both questions is 'yes' a decision could qualify as automated because it would be made in the absence of other relevant factors that could imply a decision different from that made by an AI system.

Our double test for 'reasoned automated reliance' will be mentioned further in this study. Presumably, integrating such a test in the AILD/R-PLD framework would reveal a can of worms that neither the EU legislature nor the CJEU are keen on opening. Indeed, to inquire if a human agent would have made the same decision as an AI system in a given circumstance presupposes that there be a standard (say, a variant of the reasonable person test) serving as referent for the assessment of this type of *ex hypothesi* reasoning. The discussion on the possibility for such a test to emerge is beyond the scope of this paper and will, no doubt, be developed in a future study. May it suffice stressing at this stage that, if 'automated decision' within the meaning of Article 22 GDPR means *automatic reliance on AI output* (slavish or reasoned) the *effectiveness* of the right to explanation would depend on a data subject's *ability to prove* and *explain* that reliance. If the data subject fails to do so, they might not be able to exercise the right to explanation because the decision at stake would not be considered as automated.

---

426 CJEU (Opinion - AG Pikamäe), 16 March 2023, *Schufa Holding et al*., case C-634/21, EU:C:2023:220.

427 CJEU, 7 December 2023, *Schufa Holding et al*., case C-634/21, EU:C:2023:957.

428 *Id*., pt 47.

Our goal here is not to suggest a 'new' normative interpretation of Article 22 GDPR. In the future, both the CJEU and scholarship will no doubt enlighten us more on what 'automated decisions' are in connection to the GDPR. With explanatory accuracy as *fil rouge* of this paper, our brief comment on said Article 'merely' serves the purpose of canvassing the key features expected from explanations in the context of automated data processing. The feature to keep in mind for the remainder of this article is - again - the *holistic* nature of explanations: these should concern *all the stages* of a given data processing and deliver *ad hoc* and *post hoc* understanding to the data subject.

Assuming that the GDPR is a useful referent for the explanations provided under the EU's AI regulation (AI Act, AILD and R-PLD), the victims of harm associated with high-risk AI systems should be entitled to request explanations on the transparency/explainability constraints embedded in the system (*ad hoc* explainability) as well as on the concrete unfolding of a given decisional process (*ad post* explainability). However, the procedural EU regulation of AI creates systems of evidence that only support *ad hoc* explainability. What matters is that the human agents (programmers, users, deployers, importers etc) be able to explain that they did all they could to create well-performing (transparent, robust, explainable etc) AI technologies. These are no doubt important explanations. But shouldn't the victims be the ones to decide what they need to know? If the cited North-Amercian caselaw shows us anything, it is that litigants do have the tendency to require *post hoc* explanations that is, information on how an AI system actually arrived at a decision *in concreto* (*i.e.* in their particular case). Under the relevant EU instruments, it is not a given that the disclosure of such information will be authorized, because victims are restricted as regards the *types of evidence* they can ask to have access to. As will be argued, the 'holistic' concept of explanation the GDPR seems to warrant is imperfectly (because partially) translated in AI-specific instruments like the AILD.

C. 'It's about Understanding if (Technical Standards were Observed)' - Said No One… Except the EU Legislature

A paradox characterizes the EU' forthcoming regulation of AI liability that is, the AILD and R-PLD. On the one hand, we observe *openness*: both instruments 'open up' a procedural pathway for victims of harm through the right to request disclosure of evidence. Ideally, this right is meant to provide victims with the understanding necessary for them to establish and explain the causal link between an AI system and a harm suffered, thus increasing their chances of justifying compensation. On the other hand however, we detect a *restriction*: the evidence that victims can request disclosure of is quite limited in scope. Indeed, if disclosed, that evidence can only support *ad hoc* explainability, providing understanding on whether *a priori* technical standards were complied with. When exercised, the right to request disclosure does not make available any meaningful or relevant information on a system's functionalities or decision-making processes having actually resulted in the suffering of harm (*post hoc* explainability).

The limitation to *ad hoc* explainability is, no doubt, useful because, by virtue of the cited instruments' provisions, that explainability calls for evidence based that the EU legislature deems as necessary to presume fault or defectiveness (**Sub-Section 4.3.1.**). However, a closer look at the systems of evidence in the AILD and R-PLD reveal a series of inconsistencies, which beg the question of whether the procedural

rights these instruments laudably recognize can, in practice, be conducive to an *effective*, *truly meaningful participation* in, and *fair adjudication* of AI liability disputes (**Sub-Section 4.3.2.**).

1. The Right to Request Disclosure of Evidence

The AILD creates a fault-based system, placing on the claimant the burden to prove the defendant's fault. 'Fault' is defined as "*human act or omission* which does not meet a duty of care under Union law or national law that is directly intended to protect against the damage that occurred."[429] From the perspective of liability scholarship, this definition is unsurprising: it assumes that 'faulty' behavior is equivalent to unlawful behavior which only a human agent can be accused of.

The AILD pursues a double regulatory objective: first, it seeks to establish common rules on the *disclosure* of evidence on high-risk AI systems in view of enabling claimants to "substantiate a non-contractual fault-based civil law claim for damages."[430] Second, it regulates the overall "burden of proof in the case of non-contractual fault-based civil law claims brought before national courts for damages caused by an AI system."[431]

It can be argued that the right to request disclosure of evidence in the AILD gives a specific procedural expression to the right to transparency and human explanation, originally enshrined in the GDPR. In the Directive, the beneficiaries from said right are victims of harm caused by *high-risk* AI systems. That benefit is not automatic: a claimant cannot - merely - rely on their status of (alleged) victim to request that evidence be disclosed by the defendant. On the contrary, they carry the burden of proving the merits of the case by establishing that, prior to fact disclosure request brought before a court, they had undertaken all proportionate attempts to "gather the *relevant* evidence from the defendant."[432] Only when those attempts fail, may the victim go before a national court and ask that it order the disclosure requested.

When the court finds it plausible to issue such an order, the disclosure should be "necessary and proportionate," taking into consideration the legitimate interests of all parties, in particular any limitations that might stem from the protection of trade secrets within the meaning of Directive 2016/943,[433] as well as of any confidential information related to, say, public or national security. If, after the issuing of such an order, a defendant (user or provider) fails to comply, national courts shall - and here's the kicker - "*presume their non-compliance with a relevant duty of care*,"[434] this

---

429 AILD *cit. supra*, Preamble, pt 22.
430 *Id*., Art. 1(a).
431 *Id*., Art. 1(b).
432 *Id*., Art. 1(2) (emphasis added).
433 Directive 2016/943 of the European Parliament and of the Council, of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, *OJ L 157, 15.6.2016, p. 1. Pursuant to Article 2(1) of this Directive, a 'trade secret' is interpreted as information which - cumulatively - meets three requitements: it is a secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question (a); it has commercial value because it is secret (b); it has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret (c).*
434 AILD, *cit. supra*, Art. 3(5) (emphasis added).

presumption being essentially justified by another presumption that "the evidence requested was intended to prove for the purposes of the relevant claim for damages."[435]

Article 3 AILD is echoed *mutatis mutandis* in Article 8 R-PLD which also recognizes a right to request disclosure of evidence. Under the R-PLD, an injured party claiming compensation for damages caused by a defective product (such as a biased AI) may bring their disclosure request before a national court. The claimants acting under the R-PLD - much like those relying on the AILD - are required to present "facts and evidence sufficient to support the plausibility of the claim for compensation."[436] Here again, national courts are bound by a principle of proportionality and the legitimate interests of the parties[437] while being mindful of any confidentiality restraints related to, say, the possibility to disclose trade secrets.[438] If the defendant refused to comply with the order to disclose evidence, the defectiveness of the product will be presumed.[439]

Though much can be said based on the sheer comparative reading of Articles 3 AILD and 8-9 R-PLD, we will limit our comments to two key points: first, the *effectiveness* of the right to request disclosure of evidence; second - and more importantly - the *conditions for the formation* of the presumptions of fault and defectiveness.

Regarding the first point, there is little doubt that, on paper, the cited Articles are laudable. They finally recognize a procedural right to access evidence, which part of scholarship has been adamantly pleading for since the early days of AI's regulatory discourse.[440] However, the *effectiveness* with which this right will or should be exercised remains unclear, mainly because of the national courts' discretion in the instruments considered. Indeed, both the AILD and R-PLD admittedly introduce minimal harmonization, not seeking to reduce or eliminate the Member States' discretionary powers. This of course comes at the risk of enhancing the disparity regarding the conditions under which disclosure of evidence can be granted: neither the AILD nor the R-PLD offer any guarantee that, say, French and German courts when applying their respective national laws, will order said disclosure *in the same* conditions.

To illustrate this risk of disparity, consider the following automated recruitment scenario. As we have argued elsewhere[441] it follows from the CJEU's caselaw that in 'ordinary' (non-automated) recruitment cases, the recruiters are under *no obligation* to disclose information on the criteria used to select job applicants.[442] Let us then imagine an applicant who suspected biased automated recruitment, following which they decided to request, from the recruiter, information on the algorithm's functionalities as well as on the profiles of the job applicants shortlisted for an interview. Indeed, to be able to argue, say, ethnic bias, a job applicant of color would need to access the selected shortlist, whose racial background would support (or not) that applicant's suspicion of

---

435 *Id*., Art. 3(5).
436 R-PLD, *cit. supra*, Art. 8(1).
437 *Id*., Art. 8(2).
438 *Id*., Art. 8(3).
439 *Id*., Art. 9(1).
440 Ljupcho Grozdanovski, "In search of effectiveness and fairness in proving algorithmic discrimination in EU law" (2021) 58 *CMLRev.*, 99.
441 *Ibid*.
442 CJEU, 19 April 2012, *Meister*, case C-415/10, EU:C:2012:217.

being discriminated against. However, recruiters are not often keen on making transparent their candidate lists and, in EU law, they have not obligation to do so, as confirmed by the CJEU in the *Meister case.*[443]

In our automated recruitment scenario, suppose the recruiter refused to disclose the information requested, pushing the applicant to request that disclosure before a court. The court's decision could go in one of two ways. On the one hand, the national judge can refer to the CJEU's *Meister* case concluding that, under EU non-discrimination law, recruiters are, indeed, not required to share information on the conditions under which recruitments had been performed. Based on this caselaw, the court could consider that: 1. bearing in mind the exceptions listed in the AILD, it would be within the employer's *legitimate interest* not to make known the criteria and procedures they followed in selecting applicants; 2. in EU non-discrimination law, recruiters are, anyway, not bound by an obligation to disclose such information. In such circumstances, it is not unreasonable to assume that a victim's request for disclosure on the grounds of AILD/R-PLD would be rejected.

On the other hand, however, the court could refer to Annex III of the AI Act which lists access to labour as a sector where high-risk systems are used.[444] To verify if the recruiter in fact complied with the AI Act, it might order that they disclose the evidence requested by the claimant… even if this meant going against the CJEU's longstanding caselaw on the recruiters' (non-existent) obligation to share recruitment information with unsuccessful job applicants.

Considering that the AILD is not yet binding, these are of course speculative observations. But they do allow us to make an important point: national courts will be left with considerable freedom to assess the grounds on which they order (or not) disclosure of evidence, the danger being that the benefit from the *right* to request such disclosure may vary from one national law to another. In the absence of specific guidelines in the AILD, the national courts' decisions may be based on a variety of criteria, ranging from the type of evidence at stake, the national procedural and data protection requirements, EU data sharing and data protection requirements, to national or the CJEU's constant caselaw in the sector(s) concerned. The vagueness of those criteria might have the effect of not always providing claimants with the *effective* possibility to access the evidence they need to launch proceedings, which is of course alarming. What if an HR system was indeed biased, but a national court decided against ordering any disclosure of evidence relative to that system? Should we accept that, due to the differences between national procedural laws, there will be cases of AI liability that will go undetected and unsanctioned?...

Second, the presumptive mechanism in Articles 3 AILD and 8-9 R-PLD is surprising from a perspective of fairness: *the defendant's refusal to disclose information seems to be interpreted as a confession of guilt* of sorts. The reasoning

---

443 *Id*., pts 13 seq.

444 AI Act, *cit. supra*, Annex III, pt 4 (post-compromise): "AI systems intended to be used for recruitment or selection of natural persons, notably for placing targeted job advertisements, screening or filtering application, evaluating candidates in the course of interviews or tests; (b) AI systems intended to be used to make or materially influence affecting the initiation, promotion and termination of work-related contractual relationships, task allocation based on individual behavior or personal traits or characteristics, or for monitoring and evaluating performance and behavior of persons in such relations."

seems to go as follows: if the defendant did not wish to share information, it must be because they 'have something to hide' in terms of their compliance with a legally prescribed duty to care or applicable safety requirements. In other words, non-compliance with a *procedural duty* (to disclose information) constitutes the basic fact (*indicium*) that gives rise to the presumption of fault *i.e*. non-compliance with a *substantive duty* (to observe applicable technical legislation). This procedural-to-substantive leap is rather 'light': it is similar to presuming that when a person skips lunch, it is because they have an eating disorder (which might be the case, but additional evidence would be needed for this inference to hold).

The peculiarity of the presumptive reasoning in the AILD and R-PLD does not end there: when a presumption of fault or of defect is established, the claimant - we might think - is discharged from further adducing any evidence of fault or defectiveness. Interestingly, this is not the case. In the AILD, the burden of *proving* fault reappears in Article 4 relative to the *presumption of causation*.

2. The Exercise of the Right to Request Disclosure of Evidence

The 'incoherence' in the exercise of the right to request disclosure of evidence finds two main expressions. In the AILD, the evidentiary status of fault is peculiar. When a victim seeks to establish it, fault can, under certain conditions, be presumed. When the victim seeks to establish causation, they are required to give several types of evidence which include… *proof* of fault. The question then becomes the following: how can a victim establish fault when fault is presumed (*i.e*. is not based on any solid evidence of *indicia*) (**A**)?

Much like the AILD, the R-PLD has an incoherence of its own. This incoherence pertains to the proof of defectiveness. Essentially understood as a failure to meet reasonable expectations of a normal functioning of an AI system (whatever 'normal' is),' defectiveness can be presumed in the same conditions as those under which fault is presumed in the AILD (*i.e*. refusal to disclose evidence requested). This begs the following question: when we presume defectiveness under the R-PLD, do we *ipso facto* presume fault under the AILD (**B**)?

a. Fault in the AILD: a Fact First Presumed Then Proven

Article 4 AILD habilitates national courts to presume the *causal link* between the fault of the defendant and a given output (or the absence thereof) by the AI system when *three cumulative conditions* are met: the *claimant has proven the fault of the defendant*,[445] it can be considered *reasonably likely* that the fault has influenced the output produced by the AI system (or the failure to produce an output),[446] the claimant has proven that the output produced by the AI system has given rise to the harm suffered.[447] Similarly, Article 9 R-PLD (titled 'Burden of proof') states that the presumption of defectiveness is established when: 1. the claimant *proves* that a defendant refused to comply with the obligation to disclose 'relevant evidence' upon a court order;[448] 2. they establish that the product did not comply with mandatory safety

445 AILD *cit. supra*, Art. 4(1)(a).
446 *Id*., Art. 4(1)(b).
447 *Id*., Art. 4(1)(c).
448 R-PLD, *cit. supra*, Art. 9(2)(a).

requirements laid down in Union law or national law, intended to protect against the risk of ham occurring;[449] 3. they establish that the harm was caused by an *obvious malfunction* of the product during normal use or under ordinary circumstances.[450]

There is much to unpack from these provisions. Let us begin by highlighting the - intentionally? - vague wording of the AILD: how could a claimant prove the 'reasonable likelihood' that the defendant's fault was causally connected to the harmful output of a given system? From the perspective of liability doctrines, the proof needed in the context of a 'reasonable likelihood' situation would involve demonstrating that the defendant's actions played a *contributing role* in (*i.e*. was a contributing cause to) a harm materializing. Judging by the wording alone of Article 4 AILD, the standard of proof seems to be low - 'reasonable likelihood' as opposed to *conclusiveness* (in civil cases, preponderance of evidence). Bearing in mind the minimal level of harmonization stemming from the AILD, we can assume that that national courts will assess 'reasonable likelihood' in reference to the standards of evidence contained in their national laws which - as argued earlier - might differ from one Member State to another, adversely affecting the effectiveness of the claimants' procedural abilities. Setting aside the disparity between the Member States' laws of evidence, let us, in an *élan* of prospection, anticipate a claimant's explanatory and evidentiary strategy in establishing this 'reasonable likelihood' standard.

Take the following hypothetical: a biometric identification system is used by a Member State's authorities to assess asylum applications. Nationals from a specific country notice they are systematically refused asylum, pushing them to suspect that the system disregards applications submitted by citizens of that country. Suppose that they decided to launch an action of discrimination on the grounds of nationality, requesting that the competent authorities disclose information about the system's accuracy. Imagine the authorities refused, pushing the national court to presume their fault under Article 3 AILD. So far, so good: by virtue of this presumption, the victim would be discharged from their duty to establish the *cause* of their harm (*i.e.* fault). The story does not stop there, however.

Under Article 4 AILD, the victim should *further argue (and prove) causation* and *harm*. To do so, they would need to *positively prove* fault. The million-dollar question is thus the following: *what is the point of presuming fault if a victim still needs to establish it when proving causation*? In other words, how can a victim prove that the defendant's conduct 'reasonably likely' impacted a system's output, if the latter refused to disclose any relevant evidence that the victim might use to argue causation?

The fact that the claimant's burden to establish fault is not really removed in the AILD, is confirmed in Article 4(2) which goes on to specify the *relevant facts* to be established by the claimant, depending on whether the defendant is a provider or a user. When the defendant is a provider, said Article states that the conditions pertaining to the proof of causation shall be met, *only where the complainant has demonstrated* that the provider or, where relevant, the person subject to the provider's obligations, failed to comply with any of the requirements laid down in Chapters 2 and 3 of Title III of the AI Act.

---

449 *Id*., Art. 9(2)(b).
450 *Id*., Art. 9(2)(c).

The claimant is called to - somehow - give evidence that supports *ad hoc* explanations, aimed at showing that if harmful output was produced, it was essentially because an AI system was ill-designed since its inception. For example, a claimant is held to present proof (and explanation) that an AI system was not developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in Article 10 (2-4) AI Act;[451] that the system "was not designed and developed" in a way that meets the transparency requirements laid down in Article 13 AI Act;[452] that it did not allow for an effective oversight by natural persons during the period in which it was in use pursuant to Article 14 of the AI Act,[453] and that it did not achieve an appropriate level of accuracy, robustness and cybersecurity pursuant to Article 15 and Article 16, point (a), of the AI Act.[454] The claimant may also establish that the necessary corrective actions were not immediately taken to bring the AI system in conformity with the obligations laid down in Title III, Chapter 2 of the AI Act or to withdraw or recall the system, as appropriate, pursuant to Article 16, point (g), and Article 21 of the AI Act.[455]

Alternatively, when the defendant is a user of an AI system, causation will be *presumed* if the claimant managed to prove that their adversary did not comply with their obligations to use or monitor the AI system in accordance with the accompanying instructions of use or, where appropriate, suspend or interrupt its use pursuant to Article 29 of the AI Act,[456] exposed the AI system to input data under its control which is not relevant in view of the system's intended purpose pursuant to Article 29(3) of the AI Act.[457]

The design of the burden for claimants in the AILD is peculiar. It allows for fault to be presumed while also requiring proof thereof so that causation can be presumed. The practical difficulty which ensues is the that of a litigant being unable to give evidence of the defendant's fault, in cases where fault was presumed *precisely because* the defendant refused to disclose evidence. It will be interesting to see how the Member States' and EU courts will deal with what appears to be a congenital incoherence of the AILD's system of evidence.

The EU legislator did foresee two circumstances where the claimants should not struggle as much for the presumption of causation to be established. First, the scenario where evidence is available, despite the defendant's refusal to give access to relevant information. Article 4(4) AILD states that, for high-risk systems, a national court shall not presume causation in cases where "the defendant demonstrates that *sufficient evidence and expertise is reasonably accessible* for the claimant to prove the causal link."[458] Presumably, this Article's refers to expert evidence similar to that used in cases like *Pickett*. To refer to our biometric identification hypothetical: the claimant could establish causation if they had access to publicly available expert reports confirming that the system used to vet asylum applications was notoriously biased. Article 4(4) AILD may be applied in line with the *factum* to *fama* shift, we discussed

451 AILD, *cit. supra*, Art. 4(2)(a).
452 *Id*., Art. 4(2)((b).
453 *Id*., Art. 4(2)(c).
454 *Id*., Art. 4(2)(d).
455 *Id*., Art. 4(2)(e).
456 *Id*., Art. 4(3)(a).
457 *Id*., Art. 4(3)(b).
458 Emphasis added.

earlier in this paper:[459] if they cannot access case- and system-specific evidence (and explanation) of causation, they could *faute de mieux* refer to *general* expert opinions which may confirm, or not the plausibility of that causation.

The second exception to the presumption of causation concerns cases dealing with systems that are not high-risk. For those, the presumption of causation shall only apply where national courts find "it *excessively difficult* for the claimant to prove the causal link."[460] Pity that the excessive difficulty exception is limited to non-high-risk systems only…

Finally, when a claim for damages is brought against a defendant who used an AI system in the courts of personal, non-professional activity, the presumption of causality shall apply only where "the defendant *materially inferred* with the conditions of the operation of the AI system if the defendant was required and able to determine the conditions of operation of the AI system and failed to do so."[461]

b. Presuming Defectiveness (Ergo Fault?) in the R-PLD

i. Defining Defectiveness: the Ambiguity of the 'Expectations of Safety'

Neither the PLD nor the revised version thereof (R-PLD) include a system of evidence organized around the notion of fault. As already mentioned, the relevant fact (*probandum*) in this instrument is *defect*, the presence of which is - in principle - independent from the manufacturer's intentional or unintentional failure to meet a legal standard of product safety.

In this context, Article 6 of the 'original' PLD defines *defectiveness* in reference to the *level of safety* consumers are entitled to expect from a product. This expectation may pertain to the presentation of the product,[462] its reasonably expected use[463] and the time when the product was put into circulation.[464] The R-PLD is slightly more elaborate on the definition of defectivenes. In the amended version of Article 6, the key referent continues to be the level of expectation of safety; however, in addition to the presentation/use/time of market placement triptych (inherited from the 'original' PDL), R-PLD includes other grounds for safety expectations which can be clustered into two families: 1. the security precautions that the manufacturer has control over and 2. the security precautions that can be 'reasonably' expected to be taken by the users.

The security precautions falling within the scope of the manufacturer's control are those that pertain to the disclosure under a "technical standardization legislation" (like the AI Act). The requirements found in this 'family' include the instructions for installation, use and maintenance;[465] where the manufacturer retains control over the product after the moment it was placed in the market, the moment in time when the

459 See *supra*, Sub-Section 2.2.2.
460 AILD, *cit. supra*, Art. 4(5) (emphasis added).
461 *Id*., Art. 4(7) (emphasis added).
462 Directive 85/374 (PLD), *cit. supra*, Art. 6(1)(a).
463 *Id*., Art. 6(1)(b).
464 *Id*., Art. 6(1)(c).
465 R-PLD, *cit. supra*, Art. 6(1)(a).

product left the control of the manufacturer;[466] product safety requirements, including safety-relevant cybersecurity requirements[467] and any intervention by a regulatory authority or by an economic operator referred to in Article 7 relating to product safety.[468]

Regarding the security precautions taken by the users, they are defined in reference to the *reasonably foreseeable use* and *misuse* of a given product;[469] the effect on the product of any ability to continue to learn after deployment;[470] the effect on the product of other products that can reasonably be expected to be used together with the product; [471] the specific expectations of the end-users for whom the product is intended.[472]

The requirements included in both families of safety expectations essentially aim at *elucidating the origin of defectiveness*. Much like the criteria for explanatory 'goodness,' defectiveness under the R-PLD is assessed against *objective criteria* (compliance with technical standards) and *subjective* ones (consumers' expectations of safety). The latter are evidentially tricky. To argue that a product had failed to meet safety expectations is to, essentially, prove a perceptible and verifiable deviation from that product's normal or intended use. Though the 'normalcy' and 'intentionality' of that use varies from case to case, the CJEU seems to - usually - consider the level of safety that a product warrants *generally* and the level of safety that consumers expect *in a specific case*. The *Boston Scientific*[473] case provides an interesting example here.

A US manufacturer of pacemakers and cardioverter defibrillators imported and marketed its products in Germany. A quality control performed after those products were released in the German market revealed the risk of premature battery depletion, resulting in loss of telemetry and/or loss of pacing output "without warning."[474] Pacemakers already used on patients were promptly replaced. However, a German insurance company assigned Boston Scientific before the German courts, requesting the payment of compensation in respect of the costs related to the implantation of the potentially defective devices. The German judges submitted questions for a preliminary ruling to the CJEU, asking if a defect could be considered as established under Article 6 PLD, if a group of products presented - merely - a risk of defectiveness (*i.e.* the defect has not yet materialized). In its response, the CJEU confirmed that the level of safety that a consumer is entitled to 'reasonably expect' is a key referent for the assessment of defectiveness.[475] With regard to medical devices, the Court stressed that "in light of their function and the particularly vulnerable situation of patients using such devices, the safety requirements for those devices which such patients are entitled to expect are particularly high."[476] Against the backdrop of this high level of expected safety, the CJEU concluded that, when there is evidence showing that a group of products *may be*

---

466 *Id*., Art. 6(1)(e).
467 *Id*., Art. 6(1)(f).
468 *Id*., Art. 6(1)(g).
469 *Id*., Art. 6(1)(b) (emphasis added).
470 *Id*., Art. 6(1)(c).
471 *Id*., Art. 6(1)(d).
472 *Id*., Art. 6(1)(h) (emphasis added).
473 CJEU, 5 March 2015, *Boston Scientific*, joined cases C-503/13 and C-504/13, EU:C:2014:2306.
474 *Id*., pt 14.
475 *Id*., pt 37.
476 *Id*., pt 39.

defective, "it is possible to classify as defective *all the products in that group or series*, *without there being any need to show that the product in question is defective*."[477]

The CJEU's ruling in *Boston Scientific* is noteworthy: the defect at issue in this case was considered proven, *based on the risk* that a group of products *might share* (as opposed to 'do share') the same defect. The Court thus recognized that there *may be a discharge* from the duty to adduce positive evidence in the presence of a strong enough presumption of defectiveness. The 'strength' of this presumption seems to be function of the type of product (pacemakers), the market in which that product is used (medical devices) and the expectations that consumers normally have in that market (high level of safety).

Assuming that *Boston Scientific* is a useful referent for the future application of the R-PLD, one cannot help but wonder if the CJEU would rely on a similar presumption of defect if it had to adjudicate a case like, say, *Loomis*? Would the Court consider COMPAS defective because of the risk - highlighted in several studies - of that system developing a bias? Intuitively, applying the *Boston Scientific* logic in *Loomis* would be an overstretch: the fact that COMPAS *may* express a bias does not mean that it will… But this was exactly what the Court ruled in *Boston Scientific*.

In principle, the discovery of a *high probability* for a defect in one pacemaker does not strongly warrant the belief that all pacemakers of a series share the same level of risk of defectiveness. Of course, the devices in *Boston Scientific* were not intelligent, performing personalized blood-pumping based on a patient's individual health chart. They were *automated*, manufactured according to standardized procedures and essentially performing the same function. The presumption of defectiveness in the cited case seems to stem from a logic that roughly goes as follows: 1. *in principle*, safe pacemakers are manufactured following rigorous protocols and high safety standards; 2. the risk of defect in one pacemaker is likely due to non-compliance with those protocols and standards; 3. it is likely that this non-compliance characterized the manufacturing of all the pacemakers in the same series; 4. a cost-benefit reasoning also shows that it is less costly to withdraw, from the market, the pacemakers from that series; 5. in light of these premises, it *may be* presumed that an entire series of pacemakers shares the same level of risk of defectiveness. Presented in this way, the CJEU's premise-to-presumption leap in *Boston Scientific* is not perfect but at least seems plausible. This plausibility is essentially warranted by the fact that pacemakers' operating and use are automated (as opposed to intelligent), which means that they present a certain level of *predictability*.

There is some doubt on whether the presumptive reasoning in *Boston Scientific* - as we presented it - can apply to high-risk AI systems for the simple reason that these can be technical standard conforming *and still be unpredictable*. A biometric-identification system performs one key function *i.e.* identification of individuals. However, the variables it might rely on for that purpose might be outside any reasonable (human) foresight. While a system may be trained in scrupulous observation of applicable technical standards, its outputs may vary depending on the contexts in which it operates. If the same system was used, by public authorities, in the screening of asylum seekers and in crime-preventing public surveillance, in the former scenario, the system may express, say, a racial bias whereas in the latter scenario, it may be perfectly

477 *Id*., pt 41 (emphasis added).

bias-free or express another bias (like gender or age). In other words, in the case of pacemakers, the *proof of a probable defect* (premature battery depletion) renders the risk of harm somewhat *predictable* and *verifiable*. In the case of a biometric identification system (or any high-risk system for that matter), the same level of predictability/verifiability cannot be applied.

Considering that the unprovability of defectiveness entails the unpredictability of AI systems' performance, the regulatory reflex in the EU was to reinforce *a priori* technical standardization in view of releasing, in the market, systems that can be plausibly - though not definitely - predictable. A term often used in the EU's regulatory jargon as referent for what might be a tolerable level of (un)predictability is the 'reasonably expected use' and 'misuse' of AI.

The European Parliament's (EP) Resolution on civil liability rules for AI, defined the notion of 'high risk' as a significant potential in an autonomously operating AI-system to "cause harm or damage to one or more persons in a manner that is random and goes beyond *what can be reasonably expected*; the significance of the potential depends on the interplay between the severity of possible harm or damage, the degree of autonomy of decision-making, the likelihood that the risks materializes and the manner and the context in which the AI system is being used."[478] For the EP, high-risk is synonymous with unpredictability ('is random and goes beyond what can be reasonably expected'). It is also an issue of degree ('significance and potential'). Intolerable levels of unpredictability are measured against several probabilities: the severity of the harm (provided it can be foreseen), the degree of autonomy and the likelihood of a risk materializing. These are, of course, general evidentiary guidelines, the concrete meaning and application of which being no doubt determined on a case-by-case basis.

The AI Act, focused on prevention of harm, mentions the *reasonably foreseeable misuse* of AI, defined as the use of a system in a way that is not "in accordance with its *intended purpose*, but which may result from *reasonably foreseeable human behavior* or *interaction* with other systems."[479] This instrument thus assumes two things: 1. that a system has a known or knowable ('intended') purpose, generating an expectation that it should operate in accordance with that purpose (e.g. recruiting workers on the basis of skill alone); 2. in light of that purpose, the system warrants a reasonably foreseeable human conduct. Both factors essentially tie into a standard understanding of human control and oversight: a predictable AI system is one that remains *within the scope of the purpose defined or intended and the risks foreseen by a human agent* (programmer or user). This observation is supported by the reading of the AI Act's provisions on risk detection and management. The risk management systems consist in integrative processes that run through the entire lifecycle of those system, and which may entail regular systematic updating. These systems include the identification and analysis of any *known* and *foreseeable* risks associated with high-risk systems;[480] estimation and evaluation of the risks that may emerge when those systems are used in accordance with their intended purpose and under conditions of "reasonably

---

478 European Parliament Resolution of 20 October 20202 with recommendations to the Commission on a civil liability for Artificial Intelligence (2020/2014(INL), *OJ C 404, 6.10.2021, p. 107,* Art. 3 of the proposed Regulation.

479 AI Act, *cit. supra*, Art. 3(13) (emphasis added).

480 *Id*., Art. 9 (2)(a).

foreseeable misuse,"[481] evaluation of other possibly arising risks based on the analysis of the data gathered from the post-market monitoring system[482] and the adoption of suitable risk management measures.[483] The risk management measures should be such that any "residual risks" (whatever those are) associated with a hazard and overall residual risk of high-risk AI systems "is judged acceptable, provided that the high-risk AI system is used in accordance with the intended purpose or under conditions of reasonably foreseeable misuse. The residual risks shall be communicated to the used."[484]

The key takeaway from the risk identification and management systems is that the so-called high risks can never be fully eliminated, but can at least be reduced to an *acceptable level*, defined in reference to that which a human can *reasonably foresee*.[485] It remains however that human *foresight* in this context is reasonable, not panoptic: harm may occur without a human agent being able to foresee the (risk of) defect which might cause it. In light of this, the R-PLD introduces a lightening of the burden to prove defectiveness using a well-known evidentiary device used in contexts of uncertainty. Enter the presumption of defectiveness.

ii. Presuming Defectiveness

A reading of the system of evidence in the R-PLD shows a multifaceted *onus probandi*. To be entitled to compensation, Article 9(1) requires that the claimant prove the defectiveness of a given product, the damage suffered and the causal link between the two. The system of evidence in said Article does not structurally differ from that defined in Article 4 PLD.[486] The novelty in the R-PLD is that it establishes a *presumption of defectiveness* when any of the following conditions (*ergo* not all of them cumulatively) are met: 1. the defendant has failed to comply with an obligation to disclose relevant evidence at their disposal;[487] 2. the claimant establishes that the product does not comply with mandatory safety requirements laid down in Union law or national law, intended to protect against the risk of the harm suffered;[488] 3. the claimant establishes that the harm was caused by an obvious malfunction of the product during the normal use or under ordinary circumstances.[489]

In the first two cases, the normative kinship between the R-PLD and the AILD is apparent: the presumption of defectiveness seems to be formed under the *same conditions* as the presumption of fault. Like fault, defectiveness is presumed when a defendant refuses to disclose evidence requested by the claimant which brings up an interesting question: *where there is presumption of fault, is there also a presumption of defectiveness*? Imagine a case of biased automated access to social benefits (a high-risk

---

481 *Id*., Art. 9(2)(b).
482 *Id*., Art. 9(2)(c).
483 *Id*., Art. 9(2)(d).
484 *Id*., Art. 9(3).
485 This is the gist of human oversight: risks to health, safety or fundamental rights should be limited to uses in accordance with a system's intended purpose of under conditions of *reasonably foreseeable misuse*, in particular when such risks persist notwithstanding the application of other requirements set out in the AI Act. See AI Act *cit. supra*, Art. 14(2).
486 PLD, *cit. supra*, Art. 4: "The injured person shall be required to prove the damage, the defect and the causal relationship between defect and damage."
487 R-PLD, *cit. supra*, Art. 9(2)(a).
488 *Id*., Art. 9(2)(b).
489 *Id*., Art. 9(2)(c).

sector in the AI Act[490]) was brought before a Member State's court. Suppose the social services concerned refused to disclose evidence on, say, compliance with the human oversight standard. That refusal would be a basic fact for both the presumption of fault *and* the presumption of defectiveness. But does this mean, in future caselaw, that the AILD and R-PLD will *apply jointly*? Only time will tell. At this stage, we can but observe that the evidentiary rationale of both instruments is the same: proof of non-compliance with technical standardization is the decisive *indicium* for both the presumption of fault and the presumption defectiveness to stand.

Second, defectiveness is presumed when the claimant shows an 'obvious malfunction of the product during the normal use or under ordinary circumstances.' Intuitively, this seems reasonable. Procedurally, it opens questions, chief among them being the proof of 'obvious malfunction.' Considering - as we did earlier - that the so-called high risks, and corresponding harms, are hardly predictable, in which circumstance would a system's malfunction be obvious? The existing caselaw shows that harm becomes manifest when it is too late *i.e*. when it had already materialized. The *Arkansas Department of Human Services v. Ledger Wood et al*.[491] case gives a good example of this.

The appellees were low-income individuals with serious physical disabilities. They were beneficiaries of a Medicaid program that provides home-based and community-based services. Registered nurses made individual assessments of the beneficiaries' needs and based on those, determined the number of hours of homecare per week. The DHS implemented a reassessment system (Resource Utilization Groups system - RUG), based solely on a set of complex computer algorithms. These algorithms took patient information gathered from 286-question ArPath assessment and placed the beneficiaries into one of twenty-three RUG tiers. It is important to stress that once a beneficiary was assigned to a tier, the nurses *had no discretion* in moving them to another tier.

It soon became apparent that the system was disastrously flawed, leaving patients without adequate care: many remained without food, in soiled clothes, were not bathed, missed key exercises, treatments and turnings, faced an increased risk of failing, became more isolated in their homes and generally suffered worsened medical conditions due to the lack of care. They brought an action under the Administrative Procedure Act (APA), arguing that the DHS did not comply with the latter. Without much difficulty, the circuit court found that the plaintiffs provided the evidence necessary to prove merits (*i.e*. the likelihood of their claims for damages being

---

490 AI Act, *cit. supra*, Annex III (post-compromise), pt 5: "(a) AI systems intended to be used by or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, including healthcare services and essential services, including but not limited to housing, electricity, heating/cooling and internet, as well as to grant, reduce, revoke, increase or reclaim such benefits and services, (b) AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud; (c) AI systems intended to be used for making decisions or materially influencing decisions on the eligibility of natural persons for health and life insurance; (d) AI systems intended to evaluate and classify emergency calls by natural persons or to be used to dispatch, or to establish priority in the dispatching of emergency firs response services, including by police and law enforcement, firefighters and medical aid, as well as of emergency healthcare patient triage system."

491 Supreme Court of Arkansas, 9 November 2017 (Opinion Delivered - Appeal from the Pulaski County Circuit Court, N° 60CV-17-442), *Arkansas Department of Human Services v. Bradley Ledger Wood et al.*, No. CV-17-183.

successful). In the appeals judgment, the appellants contested this, arguing their adversaries' failure to prove irreparable harm. Usurpingly, this argument was not found convincing. Indeed, in US caselaw, harm is 'irreparable' when it "cannot be adequately compensated by money damages or redressed in a court of law."[492] Considering the evidence adduced, the Arkansas Supreme Court found that the appellees "have provided a sufficient showing of irreparable harm to justify the circuit court's issuance of a temporary restraining order."[493]

However - and here is the interesting part - the *cause of that harm* was not the fact that the algorithm 'messed up.' It was that the DHS *made automatic reliance on its output mandatory*. This is an important point to keep in mind: the emerging caselaw shows that victims of harm are not always hostile to the use of AI systems. Their criticism is often turned toward the *level of reliance* on those systems. What they seem to look for is understanding on why a human agent presumed that an AI output was accurate and therefore reliable. Based on the explanation received (or not) they then construct, as best as they can, their own causal explanations. In *Arkansas Department of Human Services v. Ledger Wood et al.* the root of the matter was not - what the AILD would define as - fault. *No one* in this case (parties, courts) felt the need to discuss if the system used complied with relevant technical legislation, the 'fault' deriving from the reliance on the system, not its non-compliance with manufacturing standards!

With the exception of cases like *Arkansas Department of Human Services v. Ledger Wood et al.*, there will be cases (possibly the majority of them) where harm will not be as manifest. Take the topical example of a credit scoring AI: a system developed a bias against ethnic minorities, by basing its decisions namely on the applicants' places of residence.[494] Noticing - to the extent that AI can 'notice' - that credit-approved applicants historically reside in 'white areas,' the system's approval of residents in those areas was much greater than that of those living in ethnically mixed ones. In a case like this, little is self-evident both as regards the harm and the malfunction having caused it. Typically, in such a case, the best a claimant can do is *suspect* discrimination which would push them to require disclosure of evidence of that harm, allowing them to move forward with judicial proceedings.

It follows that, the systems of evidence in the AILD and R-PLD are so designed that they do not include any evidence supporting *post hoc* epxlainability. As previously mentioned, this is due to the fact that both instruments are procedural expressions of an understandable but insufficiently justified normative belief: *lawful* conduct (*i.e.* compliance with technical standards) *cannot* be the source of harm.

The 'web of presumptions' that the AILD and R-PLD establish is indeed convenient from the perspective of procedural economy but is open to criticism from the perspective of basic procedural fairness in two regards. First, there is the issue of the 'meaningfulness' of the explanations: do the AILD and R-PLD, as currently designed, support the litigants' *meaningful participation* in the resolution of AI liability disputes? Second, there is the *equality of arms* principle. When we think about AI

---

492 *Id*., at 9.

493 *Id*., at 10.

494 See Will Douglas Heaven, "Bias isnt' the only problem with credit scores – and no, AI can't help" (2021) *MIT Tech'y Rev*., available on: https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/ (last accessed on 20 Jan. 2024).

liability, we tend to focus on the victim and their ability to prove and explain causation. However, we ought not forget the *defendants i.e.* the agents who, by virtue of the AILD and R-PLD, will be presumed responsible. They too have a right to meaningfully participate in the evidentiary debate and provide the explanations necessary to make their views known. The 'hermetic nature' of the evidence systems in the AILD and R-PLD invites various critiques in terms of fairness.

IV. CRITIQUE OF THE AILD'S AND R-PLD'S EVIDENTIARY HERMETISM

To sketch out ways in which - what we call - the evidentiary hermetism of the AILD and R-PLD can be 'relaxed,' let us revisit the idea of explanatory facticity:[495] explanations, including causal ones, are fact- and context bound. Let us also recall that liability law is, in essence, a *corpus* of rules and principles that crystalized *in practice first*: presumably, people dealt with causal problems long before codified law came along to instruct litigants and courts on how to address those problems. In other words - and as already stressed - causal explanations aim at accuracy (and require evidence) so that the (fair) resolution of a dispute *can be informed*. If factual accuracy were not a prerequisite for procedural fairness, we might readily consider resolving disputes through the simple act of coin tossing.

The word of advice for the future application of the AILD and R-PLD is: *presume less, prove more, and more effectively*. In this perspective, we hinted in the Introductory portion of this paper,[496] shift from a *law-based* to a *needs-based* approach, in an attempt to 'reconnect' said instruments with the procedural needs of litigants. In this context, and based on the relevant caselaw in AI liability, one point seems beyond doubt: *post hoc explainability matters* and is even paramount for the evidence and explanations given by victims of AI-related harm (**Sub-Section 5.1.**).

As for defendants, they too should benefit from the procedural ability to receive *post hoc* explanations on a system's decisional processes. This is relevant in cases where harm occurs without the defendant having intended it, or without them having been directly involved in its occurrence. The ability to request access to evidence should - for the sake of the equality of arms principle - extend to defendants as well (**Sub-Section 5.2.**).

A. The Explanations Claimants Need: Not on Compliance with the Law, But on the Accuracy and Trustworthiness of Harmful AI Output

Bearing in mind the presumptive mechanisms enshrined in both the AILD and R-PLD, it is safe to assume that the evidentiary debates which will unfold under those instruments will largely focus on the compliance or non-compliance with the AI Act (*ad hoc* explanations). This 'straightjacketing' the debate on evidence by designating the relevant cause-harm interrelationship is a textbook example of what we earlier called *underdeterministic causal labelling*.[497] The downside is, of course, that such labelling narrows the scope of the discovery of relevant evidentiary facts, restricting the

---

495 See *supra*, Sub-Section 2.1.1.
496 See *supra*, 1 - Introduction.
497 See *supra*, Sub-Section 2.2.1.

litigants' procedural ability to give evidence and explanation *other than* that required by law. When the law declares (labels) a causal truth, it usually is dismissive of the discovery of different 'truth(s),' even if they are perhaps more accurate representations of reality than that retained by a providential legislator. This is the gist of Spinoza's 'refuge of ignorance' metaphor: a causal explanation viewed as *normative* or *nomic* will, however logically 'thin,' always trump any attempt to question its truth from the vantage point of reality. Does this mean that the EU legislature prefers the convenience of *ad hoc* explainability over the fact-accuracy that *post-hoc* explainability has the potential to provide?

Take the topical example of biased AI. In a 'wrongfulness' scenario, the parties would seek to determine if a system's output to, say, approve loans to white applicants only was due to a bias already present in the system's training data or was one the system autonomously developed. With the but-for test in mind, the question that the victim would seek to answer by giving evidence (and corresponding explanations) would be the following: "*had the system not used as criterion the applicants' place of residence, would the credit-approved applicants be the same*?"

To answer this question, they would necessarily require both *ad hoc* and *post hoc* explanations in order to have a plausible (or at least, plausibly correct) idea of what actually caused the bias. Presumably, no such debate will unfold under the AILD and R-PLD: by prescribing *unlawfulness* as a 'necessary and sufficient cause'[498] of harm, both Directives conveniently circumvent any meaningful discussion on a system's *in concreto* functioning (that is, its functioning *at the time* when the harm materialized). In short, they do seem to create a 'refuge of ignorance' in the sense that uncovering factual (causal) accuracy does not seem to be their primary concern. The AILD and the R-PLD do not offer litigants the procedural possibility to prove wrongful conduct *other than unlawfulness*. A provider's record keeping might be enlightening on the data they used to program a system but may not uncover the system's specific variable-association having resulted in, say, ethnic minorities being labelled as less likely to finish college or even get into one. *That* association is the actual cause of ethnic discrimination! Not the provider's failure to neatly keep records.

Is *post-hoc* explainability necessary at all under the AILD and R-PLD? Suppose in an 'algorithmic discrimination' scenario, experts managed to reverse-engineer biased AI output, identifying the stage in a system's decisional process where the 'glitch' happened. What would be the added value of that information for the claimant? Presumably none, in the current regulatory landscape in the EU. Neither the AILD nor the PLD give the possibility of proving machine-learnt bias through evidence showing that *no human* could be reasonably associated with a case of algorithmic discrimination.

Bearing in mind our analysis of explanatory epistemology,[499] the relevant question is the following: would the claimants need to understand how a system worked and if so, should the systems of evidence in the AILD and R-PLD include ex post explainability? For the purpose of providing fact-based causal explanations, the answer is 'yes.'

---

498 The concept of necessary and sufficient cause was discussed *supra*, 1 - Introduction.

499 See *supra*, Section 2.

Moving forward, the EU legislature and courts should probably relax their obsession with the proof of unlawfulness (*i.e*. non-compliance) and focus instead on *what litigants require* in terms of evidence and evidentiary explanations. The primary justification for this is the trend becoming apparent in the emerging caselaw on AI liability: it is not about proving (human) compliance with the law, it is about *giving reasons for (human) reliance* on harm-causing (because inaccurate) AI output. Indeed, whatever the sector concerned (tax fraud, medical misdiagnosis, [500] judicial functioning[501]) litigants look to uncover and discuss the rationales of *two interrelated decisions*: *that of the AI* and *that of the human having chosen to rely on the AI*. Explanations pertaining to AI decisions address the following question: *are there reasons justifying the belief that a system's output is accurate*? The answer to this question necessarily calls for *post hoc* explanations, delivered - as confirmed by the caselaw cited in this paper - by any means available: reverse engineering, local explainability, general explainability, general expertise on a system's accuracy…

Regarding the second (human) decision calling for explanations, the relevant question is the following: *are there reasons to justify a human agent's reliance on a given AI ouput*? To answer this question, courts tend to look at human conduct, both *ad hoc* and *post hoc*. *Ad hoc* explanations - as mentioned earlier - provide information on the (legal) standards and duties imposed on human agents in view of increasing the *trustworthiness* of a system. *Post hoc* explanations provide information on an agent's reasons to consider a system trustworthy and reliable, once output is produced.

The *Loomis* case[502] gives a good example on the necessity for both *ad hoc* and *post hoc* explanations, not only because causal explanatory epistemology requires this, but because what is at stake is the exercise of a constitutional right *i.e*. the right to be presumed innocent and not be sentenced wrongfully or based on inaccurate information.[503] Indeed, the defendant in *Loomis* contended that, unless he could review how factors were weighed and risks scored, "the accuracy of the COMPAS assessment cannot be verified."[504] He further argued that "even if statistical generalizations based on gender are accurate, *they are not necessarily constitutional*."[505]

The defendant's argument in *Loomis* is interesting: his first line of defense was to say that COMPAS's decision was inaccurate, since there was no evidence to show otherwise, *in his specific case*. It is, however, his ancillary argument that is more compelling: *even if* the decisions were found to be accurate, their application should be viewed as *unconstitutional* since reliance on those decisions would violate a fundamental right. The implication in *Loomis* is that AI output should *always* be subject to some form of *ex post* control and oversight, as well as to a comprehensive statement of reasons explaining why a human agent considered that the output was trustworthy and reliable.

---

500 See Supreme Court of Arkansas, 9 November 2017 (Opinion Delivered - Appeal from the Pulaski County Circuit Court, N° 60CV-17-442), *Arkansas Department of Human Services v. Bradley Ledger Wood et al.*, No. CV-17-183.

501 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra.*

502 *Ibid*.

503 *Id*., pt 34.

504 *Id*., pt 53.

505 *Id*., pt 79 (emphasis added).

It is also interesting to note that in *Loomis*, neither the sentencing court, nor the Minnesota Supreme court appeared hostile to the courts' use of COMPAS. On the contrary, the sentencing court's stance was that the risk assessment performed by that system could be used as a relevant factor for (1) diverting low-risk prison bound offenders to a non-prison alternative; (2) assessing whether an offender can be supervised safely and effectively in the community; (3) imposing terms and conditions of probation, supervision, and responses to violations.[506] In this context, the sentencing court considered that risk assessment performed by COMPAS may be used to "*enhance a judge's evaluation*, weighing, and application of the other sentencing evidence in the formulation of an individualized sentencing program appropriate for each defendant."[507] However - the court cautioned - the use of a COMPAS *must be subject to limitations*.[508] Risk- and needs-assessment information should be "used in the sentencing decision to inform public safety considerations related to offender risk reduction and management. It *should not be used as an aggravating or mitigating factor in determining the severity of an offender's sanction*."[509] The court's ruling on this point is enlightening in its suggestion to distinguish between (human) *decisions* and *decisive factors* for those decisions. AI systems are decision-supporting tools, not decision-makings entities! Even when they are assumed to be accurate, decision-making power should never be fully delegated to them. In many ways, their output can be assimilated to 'standard' expertise: as any type of expert evidence, AI output should be informative, relevant, support informed decisions, but never replace human decision-making power. If a human chose to base their decisions on AI output alone, *Loomis* tells us that they would need to *give reasons* on why that choice was justified.

An emerging assessment standard of the justification of human reliance on AI is a hypothetical counterfactual test which answered the following question: *what content would a human decision have, had it not involved AI use?* This is, in essence, a question the Minnesota Supreme Court sought to answer in *Loomis*, ultimately finding that even without the use of COMPAS, the circuit court would have imposed "the exact same sentence" on the defendant. As mentioned earlier,[510] this is a counterfactual reasoning typical of the but-for test. However, the risk with such a reasoning is that it might be overly hypothetical. There is a fine line between *hypothesizing* and *presuming*[511] how a human agent would have acted, without an AI system being included in the decisional process. Elucidating the exact impact an AI had on a human decision is a complex issue, deserving of a separate study. For the purpose of this paper, may it suffice stressing that *Loomis* is perhaps foretelling of what we qualified as a *needs-based explanatory approach* to AI liability. This approach consists in

506 *Id*., pt 88.
507 *Id*., pt 92 (emphasis added).
508 *Ibid*.
509 *Ibid* (emphasis added).
510 *Ibid*.
511 The difference between a hypothesis and a presumption resides in their evidentiary status and the 'strength' of the inference each presuppose. We argued elsewhere that presumptions are (indirect) evidence, the object of which are facts which, in a normal state of affairs, appear to be a probable and a plausible substitute for a fact for which direct proof is sought, but is unavailable or difficult to adduce. For presumptive inferences to hold, they require probing evidence of *indicia* (basic facts) that support the strength (and truth value) of the presumptive inference. Unlike presumptions, hypothesis do not have the status of evidence. They pertain to *possible* states of affairs which, not needing to play the role of evidence, do not need to respond to evidentiary standards like those that *indicia* must meet, in connection to presumptions. See Ljupcho Grozdanovski, « Le Probable, le plausible et le vrai. Contribution à la théorie Générale de la présomption en droit » (2020) 84-1 *RIEJ*, 39, at 71.

providing evidence and explanations on *why* there are reasons to believe that a given AI output was accurate and why the reliance on that output was justified.

This accuracy/reliance schema is not only becoming visible in cases dealing with COMPAS, but can be also seen in disputes involving other AI systems. For example, in *Cahoo v. Fast*,[512] Michigan's Unemployment Insurance agency (UIA) had used a system to detect and punish individuals having submitted fraudulent unemployment insurance claims. The plaintiffs contented that UIA detected fraud where none existed and sent little or no notice to the plaintiffs, precluding them from launching administrative appeals in the authorized delays (30 days after receiving notice). In its defense, UIA gave a *negative evidence argument*, stating that the plaintiffs had failed to demonstrate injury-in-fact because their claims were not entirely adjudicated by the Michigan Integrated Data Automated System (MiDAS).

Indeed, MiDAS performed so-called *auto-adjudication* - a process beginning with the automated generation of a flag, resulting in the automated generation of questionnaires. It then created determination based on logic trees, followed by a notice of fraud, eventually conducive to collection of taxes due.[513] Admittedly, MiDAS is not a "marvel of artificial intelligence"[514] given that a human could perform any of those activities, except the generation of the fraud questionnaire.[515] Once a default fraud determination had been made, MiDAS automatically issued three notices: 1. a primary notice of determination which confirmed overpayment from the UIA, without providing any explanation on the reasons underlying that decision;[516] 2. another notice of determination which generally informed the claimant that their actions "misled or concealed information to obtain benefits and announced that benefits were terminated on any active claims;[517] 3. a list of overpayments, accompanied by a statutory penalty for fraudulent misrepresentation of two-to-four times the amount of overpayments.[518] MiDAS made a number of errors. One of the plaintiffs in *Cahoo* argued that she had been unaware of the fraud determination and did not learn about it until she had filed for bankruptcy 'months later' (even though, she admitted to not closely following the electronic communication sent to her by the Michigan social services).

Interestingly, like in *Loomis*, the litigants in *Cahoo* presented their grievances along two lines of reasoning. First came their arguments on MiDAS' *inaccuracy*, the allegation being that the fraud determinations were "wrongfully adjudicated based on MiDAS's rigid application of the UIA's logic trees, which led to 'automated' decisions."[519] Then came the *unjustified reliance argument*: like in *Loomis*, the plaintiffs in *Cahoo* contended that UIA had wrongfully relied on the output produced by MiDAS.

Unlike *Loomis* however, in *Cahoo*, the evidentiary debate on causation was slightly different: the court did not require a *post-hoc* explanation on MiDAS'

---

512 US District Court (Eastern District of Michigan – Southern Division), *Cahoo et al. v. Fast Enterprises et al*., case n° 17-10657.
513 *Id*., at 3.
514 *Ibid.*
515 *Id*., at 3.
516 *Id*., 4.
517 *Id*., at 4.
518 *Ibid*.
519 *Id*., at 19.

(in)accuracy. It found that the plaintiffs had sufficiently demonstrated an injury-in-fact stemming from MiDAS's rigid application of logic trees, coupled with inadequate notice procedures that are "fairly traceable" to FAST's and CSG's conduct."[520] The court's operative assumption seems to have been that *the proof of harm was, itself, proof that the AI output was inaccurate*.

Cahoo marks a teachable moment for our prospections on future AI liability cases in the EU. First, it bears repeating that the assumption in *Cahoo* is that it is AI inaccuracy that causes harm, not *non-compliance with technical standards*. Based on the elements of fact (absence of notice and of explanations on the reasons for tax fraud, violation of the right to property), it was apparent that MiDAS did not perform well, rendering plausible the assumption that harm was, indeed, the consequence of *inaccurate* output (again, viewed as a wrongful act of the system, not an unlawful act of its programmer).

Second, and based on that assumption, the evidentiary debate in *Cahoo* focused on the *allocation of liability* as the court sought to identify the agent who could be plausibly seen as responsible for MiDAS's inadequate functioning. Two candidates were considered: the provider and the user. To determine which of the two was the culprit, the court applied the '*fairly traceable*' test[521] used - as the but-for test and its variants[522] - to infer, from the evidence available, the agent who should bear the responsibility of compensating harm.

In the "nebulous land of 'fairly traceable' where "causation means more than speculative but less than but-for."[523] The allegation was, essentially, that UIA's system functioned the way it did because of its *provider's* injurious actions.[524] In an attempt to shield itself from liability, the latter asserted it merely followed the State's instructions.[525] The key criterion for identifying the liable party then became an agent's *level of discretion* and *intentionality* in the programming and/or use of MiDAS. Providing advice to a third party - the court stated - that voluntarily injures another "is *constitutionally insufficient* to expose one to liability, whereas actively participating in the injury is sufficient."[526] Taking into account the elements of fact, the court found that the harm was 'fairly traceable' *to both the provider and the user*.[527]

The *Cahoo* case clarifies aspects of *Loomis.* The basic evidentiary debates in both cases revolve around the *accuracy of the AI output* and *human reliance* on that output. However, each case deals with a different variant of that debate. *Loomis* is a good example of a debate focused on proving the reliance on (in)accurate AI decision of a *public* (*judicial) authority*. As already discussed, the Minnesota Supreme Court's reasoning can be criticized, namely for the application of the hypothetical sentencing test (seeking to determine the decision a court would have reached without the use of AI). Though in *Loomis*, the Supreme Court found no automatic reliance on COMPAS's

---

520 *Id*., at 27.
521 *Id*., at 21.
522 See *supra*, Sub-Section 2.2.2. (B).
523 US District Court (Eastern District of Michigan - Southern Division), *Cahoo et al. v. Fast Enterprises et al*., *cit. supra*, at 22.
524 *Ibid*.
525 *Id*., at 23.
526 *Id*., at 24 (emphasis added).
527 *Id*., at 27.

output, the evidence it considered to assess both the system's accuracy and the reasons for reliance[528] leave us wondering if the Court's level of scrutiny would have been higher, had the allegations been made against private parties or public bodies other than courts. After all, accusing a court of being a 'slave to the algorithm' would imply total delegation of the legal/judicial decision-making, which is a troubling and alarming thought.[529]

But *which test* should we use to determine if a court was justified in automatically relying on AI output? *Loomis* does not answer this question. Future caselaw - perhaps of the CJEU - will hopefully shed more light in this regard. In *Cahoo*, the violation of a fundamental right was also attributed to a public authority. However, unlike the Minnesota courts' use of COMPAS in *Loomis*, the Michigan unemployment agency in *Cahoo* played a more *active role* in shaping the use it wished to make of MiDAS.

An interesting thought comes to mind: are we witnessing the emergence of an *active human involvement* test? This test would seek to trace back an AI-related harm to an active (intentional) human act having had a decisive impact on a system's performance. The already mentioned *Coscia* case[530] is relevant here. Seeking proof of intent-to-harm (in the case of a high-speed trading algorithm capable of spoofing), the court's approach in *Coscia* is perhaps a precursor to a more generalized, future judicial practice. In essence, the court required that proof be adduced *until a human culprit could be found*. In *Coscia*, that human turned out to be the user. Indeed, similar to *Cahoo*, it was the programmers' testimonials in *Coscia* who confirmed that the user had instructed them to create a system able to make profit… Be it at the price of spoofing.

The 'active human impact/involvement' test, performed in cases of standard Business-to-Customer (B2C) or Business-to-Business (B2B) connections, is - and has been - characteristic of cases where those connections are made possible *via* online platforms. The *Force v. Facebook*[531] case gives an interesting example here. Several US citizens argued that Facebook provided Hamas (considered in the US as a terrorist organization) with a platform that enabled attacks in Israel. Facebook did not review or edit the posts made by its users. Its terms of service explicitly stated that the users owned all the content and information posted, and exercised control over how this information was shared through users' privacy and application settings.

The liability issue in this case was, of course, whether Facebook was responsible for the content published on its platform. To address this issue, the evidentiary debate focused on determining (*i.e.* proving and explaining) if Facebook was the 'publisher' or - merely - the 'speaker' of the content provided by Hamas. To this end, it was necessary to uncover *how* Facebook used its algorithms.[532]

---

528 See Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra*

529 For an analysis of the use of automation in dispute resolution, see Bastiaan van Zelst, *The end of justice(s)?: perspectives and thoughts on (regulating) automation in dispute resolution* (Eleven Int'l Publishing, 2018).

530 See US Court of Appeals for the 7th Circuit, *US* v. *Coscia*, *cit. supra*.

531 US Court of Appeals (2d Circuit), *Force v. Facebook* (2018), n° 18-397.

532 *Id*., at 22.

The plaintiffs argued that this use fell outside the scope of publishing because "the *algorithms automate Facebook's editorial decision-making*."[533] That argument did not convince the courts who asserted that 'so long as a third party willingly provides the essential published content, the interactive service provided receives full immunity regardless of the specific edit(orial) or selection process."[534] Facebook could therefore not qualify as publisher of information, but acted as mere 'speaker' of content. Though making information more available is, indeed, an essential part of traditional publishing, it does not amount to 'developing' that information as a publisher would.[535]

Even though *Cahoo*, *Coscia* and *Force v. Facebook* address legal different issues, they share the common thread of the above-mentioned accuracy/reliance evidentiary schema, as well as a test of active (intentional) involvement of a programmer or user in shaping a system's functionalities and objectives. *This* is what litigants in cases involving AI seem to *need evidence on*! To adduce that evidence, 'systems of presumptions,' such as those in the AILD and R-PLD will not cut it. Contrary to this, North American caselaw indicates that, similar to any debate involving the proof of fault, AI liability cases demand thorough fact-finding, as exemplified by trends such as *Coscia*'s 'prove until a human is identified.' This need for proper fact-finding is understandable from the standpoint of the right to a fair trial. First, for a fair adjudication, causation must, indeed, be established through fact-based explanations, ensuring compensation is awarded based on convincing information about the reality of the harm suffered. Second, fair trials maintain their 'fairness' by guaranteeing the equality of arms *for both parties*, including those presumed liable under AILD and R-PLD.

B. The Forgotten Actors in AI Liability Trials: the Rights of Defendants

According to the CJEU, the equality of arms principle is an important "corollary" the right to a fair trial.[536] In essence, this principle presupposes a level of procedural symmetry between the parties, in particular in three regards: 1. the *allocation of procedural duties* (burdens, standards of proof); 2. the *access to relevant information* and *knowledge* (in other words, evidence) able to support of their claims; 3. *equal opportunity* to make their views known and respond to the adversary's arguments. The CJEU has recognized that, in some instances, the procedural parity between the parties in a dispute may not be absolute. Admitted limitations to the right to access evidence may pertain to the content of the evidence concerned and the safeguard of constitutional

533 *Id*., at 38 (emphasis added).
534 *Id*., at 38 (emphasis added).
535 *Id*., at 49.
536 See, *inter alia*, Gen. Court, 16 July 2014, *Isotsis v. Commission*, case T-59/11, EU:T:2014:679, pt 262.

principles like the good administration (of ongoing administrative procedures or pending trials).[537]

It remains however, that save in exceptional circumstances, the parties' equal procedural footing should be observed, allowing them to benefit from the same level of - what procedural scholars have termed - *fitness to plead*.[538] The big question is, of course, if the AILD and R-PLD comply with this (constitutionally required) level of equality? To answer this question, let us bring forth the already discussed procedural postulate both instruments share: the defendant's refusal to disclose evidence in connection to their compliance with technical legislation is enough to generate a presumption of responsibility. But which evidence could they provide in order *to rebut* that presumption?

Between the AILD and the R-PLD, the former is by far the more laconic. Indeed, Article 4(7) AILD states that "the defendant shall have the right to rebut the presumption laid down in paragraph 1." This *pro forma* recognition of the right to defense points to the fact that the AILD is largely focused on regulating the burden of the claimants, though it does not pay much attention to the *feasibility* of that burden, for the reasons previously mentioned.[539] The only point where feasibility is taken into consideration is in cases dealing with the proof of causation in connection to AI systems which do not qualify as high-risk under the AI Act. For those, Art. 4(5) AILD states that said presumption shall apply only where "the national court considers it *excessively difficult* for the claimant to prove the causal link."[540]

The AILD's assumption on defendants seems to be that they, as primary bearers of the legal duty to comply with instruments like the AI Act, are *necessarily in possession of the evidence* the claimants may request access to, and that the defendants themselves might use in their defense. This of course suggests that unlike claimants, defendants cannot request that evidence be disclosed. And why would they? As argued earlier,[541] the evidentiary debates under the AILD - and by extension, the R-PLD - will revolve around *ad hoc* explainability and be limited to debates on whether the defendants complied with relevant legislations like the AI Act. The AILD appears somewhat oblivious to the *procedural needs* of the defendants, failing to consider the

---

537 The issue of the scope of the right to access evidence has, in particular, been raised in connection to the right to access documents issued by the EU institutions - namely in the context of dispute-resolution procedures - requested by third parties (*i.e.* entities not directly concerned by a disputed involving an EU institution and adjudicated on the grounds of EU law). See e.g. CJEU, 21 September 2010, *Sweden v. API and Commission et al.*, joined cases C-514/07 P, C-528/07 P and C-532/07 P, EU:C:2010:541. A journalist association based in Sweden requested the disclosure of documents relative to infringement proceedings brought by the EC against that State. The disclosure was refused, considering that the case was still pending and that the disclosure was requested by an entity that was not party to the proceedings. Analyzing the Member States' practice on the scope of the right to give generalized and unconditional public access to evidence, AG Maduro noted, in his Opinion, that not all States recognize such access, especially when the documents requested pertain to a pending case. In practice, the exercise of this right is characterized by a search for balance between ensuring the transparency of adjudicatory procedures (including the ways in which evidence is given) and the safeguard of legitimate interests (of the parties involved in the administrative or judicial procedures concerned). See *Id.*, Opinion delivered on 1 October 2009, EU:C:2009:592, para. 29.

538 See, *inter alia*, Ronnie Mackay, Warren Brookbanks, *Fitness to Plead: International and Comparative Perspectives* (OUP, 2018).

539 See our observations on the AILD, *supra*, Sub-Section 4.3.2.

540 Emphasis added.

541 See *supra*, Sub-Section 4.3.

possibility that, like claimants, they may also require a deeper understanding of the system they have used. In other words, they might also need *post hoc* explanations to exercise the right to defense. Nevertheless, given that the evidence system in the AILD does not permit the solicitation or provision of such explanations, defendants might find themselves devoid of the practical opportunity to present evidence and articulate their perspectives. This predicament arises particularly in cases where they may not comprehend the reasoning behind their system's detrimental decision-making processes.

In contrast to the AILD, the R-PLD gives a more prominent place to defendants. In its Preamble, the R-PLD stresses that the Member States' courts should presume causation where "notwithstanding the defendant's disclosure of information, it would be excessively difficult for the claimant, in light of the technical or scientific complexity of the case, to prove its defectiveness or the causal link, or both."[542] In the interest of a fair apportionment of risk - the R-PLD continues - economic operators should be exempted from liability "*if they can prove the existence of specific exonerating circumstances*."[543]

The R-PLD indeed contains several grounds for defense. As per Article 10, the defendant can escape liability if they can prove *any* of the following: 1. if they are manufacturers or importers, they should establish that they did not place the product on the market or put it into service;[544] 2. if they are distributors, they should prove that they did not make the product available on the market;[545] 3. if it is probable that the "defectiveness that caused the damage did not exist when the product was placed in the market, put into service or, in respect to a distributor, made available on the market, or that this defectiveness came into being after that moment;"[546] 4. the defectiveness is due to compliance of the product with mandatory regulations issues by public authorities;[547] 5. when the defendant is a manufacturer, "the objective state of scientific and technical knowledge at the time when the product was placed on the market, put into service or in the period in which the product was within the manufacturer's control was not such that the defectiveness could be discovered."[548] All exemptions converge in their demand for evidence of awareness (or foreseeability) regarding the risk of harm. In scenarios (1) and (2), the defendant should prove that they were not responsible for the commercialization of a 'defective' AI, arguing their *lack of relevant knowledge* on any existing or potential risks of harm. In scenario (3), the defendant should prove that the risk of harm was unforeseeable, having emerged after the system's release in the market.

---

542 R-PLD, *cit. supra*, Preamble, pt 34. The 'technical and scientific complexity' is - according to the R-PLD - a case-by-case issue and depends on various factors such as the complex nature of a product (e.g. an innovative medical device), the complex nature of the technology use (e.g. machine learning), the complex nature of the information and data to be analyzed by the claimant and the complex nature of the causal link (e.g. the link between a pharmaceutical or food product and the onset of a health condition, or a link that, in order to be prove, would require the claimant to explain the inner workings of an AI system). See *ibid*.

543 R-PLD, *cit. supra*, pt 36 (emphasis added).

544 *Id*., Art. 10(1)(a).

545 *Id*., Art. 10(1)(b).

546 *Id*., Art. 10(1)(c).

547 *Id*., Art. 10(1)(d).

548 *Id*., Art. 10(1)(e).

Scenario (4) is peculiar because it alludes to the case - not mentioned in the AILD - of *harm occurring in spite of a manufacturer's lawful conduct* (*i.e.* compliance with mandatory technical standards). By including this, the R-PLD fills a gap in the AILD regarding actions for compensation of harm occurred in the presence of evidence showing the defendant's *lawful* conduct. Here again however, the element of knowledge/foreseeability comes into play: the defendant would presumably seek to establish that their compliance with the AI Act warranted the assumption that a system was risk-free or that the technical standards followed did not allow for a risk of harm to be reasonably foreseen.

Finally, scenario (5) makes a clear allusion to expert evidence. Referring to the 'state of scientific and technical knowledge,' a defendant could escape liability by offering expertise likely to convince a court that the risk of harm was undetectable. In our opinion, and judging by the caselaw cited throughout this paper, *expert evidence* will most certainly play a prominent role in the future evidentiary debates on AI liability in the EU. In applying the R-PLD, the Member States' and Union courts will, no doubt, be called to define the probative value of the expertise brought forth by the parties. The *Pickett* and *Loomis* cases give a glimpse into a possible 'battle of experts' which will likely become exacerbated as AI technologies continue to evolve. For each expert opinion confirming the general accuracy, reliability and trustworthiness of an AI system, there will likely be a competing study arguing the contrary. We can expect to see, in the EU, the emergence of a *probative value test* which may include criteria similar to those included in the previously discussed Bradford-Hill test.[549]

In this context, one important question remains open as regards the right to effective defense: mirroring the right of claimants to request disclosure of evidence, should the grounds for defense in the R-PLD, and even the AILD, be interpreted as including a right, for defendants, to ask for independent experts, possibly for the purpose of reverse-engineering a given AI output? It is too early to tell, namely because the cited instruments are not yet binding. However, if a defendant sought to argue that a defect (like a bias) occurred *after a system had left their sphere of control*, they would naturally need to somehow prove this. The most probative evidence here would be the opening of the 'black box' which, as *Pickett* shows, can be an arduous, time-consuming process.

The deeper question is, of course, if the systems of evidence in the instruments considered should be more permissive to *post hoc* explainability, as a set of explanatory methods and techniques conducive to *understanding* of how specific systems worked (their compliance with the AI Act notwithstanding). For the sake of ensuring high levels of fairness of future AI liability cases, we might argue that *post hoc* explainability does indeed appear to be necessary, if the aim is to allow both parties to exercise their constitutional rights with equal effectiveness. Not only should claimants be able to understand the stages of causation having resulted in harm, but defendants might, depending on the facts of a case, also require such understanding: consider a recruitment algorithm displaying an unfair bias, with neither its programmer, user and potential victim having understood the reasons and methods behind the development of that bias.

---

549 See Susan Haack, "Correlation and causation. The 'Bradford Hill criteria' in epidemiological, legal and epistemological perspective," *cit. supra*.

It is yet to be seen if, when confronted with the difficult access to certain forms of *post hoc* explainability - such as reverse-engineering - the EU courts will align with their North-American counterparts, as regards the types of expert evidence they might view as admissible when direct evidence of causation is unavailable. The shift of focus to *general expertise* on specific AI systems is, as previously discussed, open to criticism: general expert opinions can support a belief in the overall trustworthiness of an AI system, but they prove nothing on that system's performance in connection to a specific harm. To resolve this conundrum, the available caselaw points to an alternative: true, general expert opinions do not establish *in concreto* (local) AI accuracy but can justify the defendant's *reasons to rely* on that system's output. The accuracy/reliance schema reappears again; we have discussed it earlier and will not revisit it here. May it suffice stressing that there is little doubt that explanations on a system's 'inner workings' are the preferred evidence, when understanding causation in AI liability cases is concerned. What litigants *need* are not statistics on Tesla cars' performance in the last five years, nor do they need to know if the manufacturing standards of Tesla cars were complied with. What they need is understanding on why *in their case*, the car made a right instead of a left turn.

However, if that type of understanding is impossible because the evidence is not accessible, the emerging caselaw reveals a shift in the explanatory enterprise from 'understanding the machine' to 'understanding the human using the machine.' The inevitability of human agency brings us back to Spinoza: considering our observations on the EU's regulation of AI liability, are we ensnared in a refuge of ignorance?

CONCLUDING REMARKS: THE AILD, THE R-PLD AND THE REFUGE OF IGNORANCE THEY BUILT

Do the AILD and the R-PLD offer a refuge of ignorance when we grapple with causal knowledge and explanation in the field of AI liability? To answer this question, we must consider the type of knowledge about facts these instruments are conducive to.

Can litigants rely on them to request the evidence and gain the understanding *they need* to causally explain the harms suffered? Alas, no. Neither of the cited instruments includes the possibility for the parties to engage in discovery proper, for the purpose of determining if a given AI-harm association was correlative or causal.

Why is it that the AILD and R-PLD fail to support *proper* discovery and explanation of causality? We have already mentioned a key component of the answer: the 'cognitive disturbance' in acquiring causal knowledge about AI lies in the potential revelation that a harm may be causally linked to an intelligent system rather than a human agent.

The AILD and the R-PLD both grapple with the current dilemma in liability doctrines, which involves choosing between liability regimes designed around *criteria for allocation of liability* and regimes designed around *criteria of discovery*. Historically, those sets of criteria were not mutually exclusive because, prior to the advent of AI, causal truths derived from discovery would reliably trace back to human culprits. However, under the influence of AI, the long-standing belief in the responsible human can be brought into question, since it no longer holds universally (*i.e.* in all cases). In spite of this, we continue to be - so to speak - *discovery-phobic*, preferring not to delve too much into facts and, with by doing so, take the risk of uncovering that

an AI system had acted without apparent human intervention. Consider the consequences of such a discovery: if evidence showed that an intelligent system caused harm *by itself*, we would need to rethink the concept of agency as cornerstone of liability in law (criminal and tort, civil and contractual).

Given our reluctance to acknowledge that AI systems can display signs of agency, we understandably cling to what we've always known to be true: only *human* agency can, directly or indirectly, be conducive to harm. To refer back to Spinoza: our preference for the human agency principle is, in many ways, not different from *choosing to believe* that stones fall from roofs because God wants them to, not because of a combination between factors like the stone's weight, the speed of the wind and gravity.

At the end of the day, the AILD and R-PLD are really not avant-garde. Consider the operative assumptions of their systems of evidence: 1. compliance with the AI Act's provisions (especially those targeting high-risk AI systems) is enough to reduce or eliminate the risks of harm; 2. if harm does ensue, it is because (and *only* because) the AI Act (or similar legislation) was not observed; 3. agents who refuse to share information on their compliance with the AI Act - in a way - confess to being at fault or to the defectiveness of a system used. Based on these assumptions, said systems of evidence are designed in such a way that, whatever evidence and explanations the parties request and give, the resulting 'knowledge' will always showcase that a human (dis)obeyed the law, rather than uncover the factors that played into an AI system acting in the way it did.

From the perspective of the epistemology of knowledge, the AILD and R-PLD are not perfect but their underlying motives are certainly understandable. The trickier question is whether their design is *procedurally fair*, from the litigants' standpoint. This entire paper is dedicated to arguing why the answer to this question is 'no.'

As mentioned earlier, procedural fairness translates - or ought to translate - to frameworks of abilities which give tangible expression to the principle of equality, namely in the ways in which litigants give and receive evidence and (causal) explanations. Ideally, the exercise of these entitlements should support the litigants' *meaningful participation*. This concept of 'meaningfulness' - *from the perspective of individuals, not legislators*! - is a recurring theme across the points raised in this paper: we contended that a crucial element in enhancing the believability of explanations lies in their level of significance to those receiving them.[550] We also argued that the meaningfulness of evidentiary debates is largely function of how effective the litigants' abilities are in accessing the evidence and giving explanations that *they consider* as relevant for the expression of their views.

Through the prism of this idea of meaningfulness - specifically referring to the litigants' 'meaningful participation' in trials - the AILD and R-PLD are open to criticism. Following up on our *needs-based approach* to AI liability, we examined what we consider to be topical examples of the emerging caselaw, revealing a trend which shows that, from the litigants' perspective, *explanations about causation do matter*. While legal compliance is important, it is the last thing litigants (and even courts) are likely to flag as a key explanatory factor in AI liability cases. As previously argued, the emphasis

[550] See *supra*, Sub-Section 2.1.2.

in what litigants 'need to understand' is underscored in two aspects: first, the accuracy of a specific AI output (requiring explanations related to *all the factors* influencing the system's output, both *ad hoc* and *post hoc*), and second, the rationale behind why human agents believe that the output was genuinely accurate and justified reliance. In essence, litigants seek to understand the rationalities involved in a case of AI use having resulted in harm: on the one hand, the *rationality behind the automated decision*, on the other hand, the *rationality behind the human decision* to rely on it. This suggests that, for the purpose of causally explaining AI-related harm, human and non-human behaviors are viewed as components of a *single causal chain*.

In summary, *proving* and *explaining causation* is crucial for the adjudication of AI Liability cases. For the sake of accuracy, meaningful participation (of litigants) and fairness (of judicial decisions), *post-hoc* explanations should be incorporated into the causal explanations and evidence presented under the forthcoming procedural regulation in the EU. The rationale behind that integration is simple: "we don't want theories. We want facts!"[551] - a statement which holds even more weight when we consider that it is evidence and *post-hoc* explanations that provide the best opportunity for dispute resolution in the field of AI liability to be informed and by that, more fair.

[551] Doris Lessing, *The Grass is Singing* (Fourth State, ed. 2013), at 22.

THE RIGHT TO A JUST REMEDY IN PRIVATE LAW—A RIGHT AND A HUMAN INSTINCT: AN ANALYSIS OF HOW GREED AND LAWLESSNESS REMOVES CONFIDENCE

Nicolas Garon*

Abstract: This article explores the fundamental concept of justice and the right to a remedy in private law. Specifically, the question of whether the belief that justice will prevail is a spoiled concept to have, and only shared by populaces in places without corruption, rule of law issues, etc. The paper delves into the historical and global recognition of the rights of obligees, and scrutinizes the diminishing trust and conviction in justice, especially in jurisdictions plagued by corruption and weak rule of law. Bribery and corruption in judicial systems undermine access to remedies, particularly for the poor, and contribute to the erosion of faith in the judiciary. The analysis extends to non-state judiciaries, exploring the role of tribal courts in addressing private law matters in regions with challenges in formal legal systems. Ultimately, the paper contends that the instinct for justice is a timeless and universal human trait, inherent in the evolution of legal systems. While corruption and external influences can erode confidence, the desire for a just remedy remains engrained.

Keywords: Justice; Corruption; Judicial Corruption; Instinctive Access to Justice; Sociology of Remedies and Their Opinions; Comparative Legal Systems

* Southern University Law Center, United States.

Table of Contents

INTRODUCTION

In private law, the rights to justice and equity—with a right to remedy as a safeguard and legitimizer of enforceability—are privileges the United States, and any jurisdiction with a rule of law, bestows upon its citizens.[1] Specifically, in the United States, a bedrock of our legal system in matters of private law is the right to a remedy and access to restorative relief.[2] Arguably, private law exists as a means to exercise such claims,[3] and is also for ameliorating individual justice.[4] Whether in negligence, intentional torts, property, or breach of contract claims, the plaintiff or obligee is owed the right to a remedy.

The purpose of contract law revolves around morality and enforcement. Arguably, contract law can be understood, in its purpose, to enforce *moral* obligations.[5] It is considerably a moral duty to uphold a contract,[6] which highlights the instinctive nature of contract law's scope. In fact, the word "moral," as it relates to some moral obligation, can often be found in contracts literature,[7] highlighting the strong interconnection between emotion, which some understand as a natural obligation between persons, and an internal belief that a person owes some level of standards to another.

One might argue, especially in a purely individualistic world, one is owed little in a public to private citizen relationship, and in turn, nothing in a person-to-person relationship.[8] Further, one may argue that individualism is on the bottom layer of this agency in regard to rights of accountability, and above it are tribalism, which gives agency to induvial, and then on top, an organized state with a just judiciary.[9]

I believe that in any setting, whether individualistic, tribal, just, authoritarian, or corrupt, a self-entitlement to justice and an owed feeling of accountability and right to remedy with a looming liability to keep things in check exist in any social structure between every person, even if systematically, a judiciary acts negligently in delivery. Regardless, if a certain government or judiciary overlooks large issues in lawlessness, private law legitimacy always remains. This idea can be found in legal systems east to west, corrupt to fair, autocratic to democratic.

---

[1] Daphna Lewinsohn-Zamir, *Do The Right Thing: Indirect Remedies in Private Law*, 94 B.U. L. REV. 55 (2014).
[2] Hugh Collins, *Private Law, Fundamental Rights, and the Rule of Law*, 121 W. VA. L. REV. 1 (2018).
[3] *See* Andrew S. Gold, *A Moral Rights Theory of Private Law*, 52 WM. & MARY L. REV. 1873 (2011).
[4] Nathan B. Oman & Jason M. Solomon, *The Supreme Court's Theory of Private Law*, 62 DUKE L.J. 1109 (2013).
[5] *See* Stephen Michael Waddams, *The Modern Role of Contract Law*, CANADIAN BUS. L.J., 1983
[6] Daniel Markovits & Emad Atiq, *Philosophy of Contract Law*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2021).
[7] CHARLES FRIED, CONTRACT AS PROMISE: A THEORY OF CONTRACTUAL OBLIGATION 7–8 (2d ed. 2015).
[8] Claude S. Fischer, *Paradoxes of American Individualism*, 23 SOCIO. F. 363 (2008).
[9] Danny Jones, *CIA Spy Explains How the United States Betrayed Him | John Kiriakou*, YOUTUBE (May 23, 2022), https://www.youtube.com/watch?v=dfYnLqYEnfw.

In contract law in Louisiana's mixed civil law system, obligees have rights, as do obligors, like the rest of the common law United States.[10] In Louisiana, as well as throughout the United States, various types of breach can be remedied.[11]

The broad idea that a liable party owes restorative retribution to a breached plaintiff can be understood through jurisprudence. For example, Louisiana has a statute stating "[e]very act whatsoever of man that causes damage to another obliges him by whose fault it happened to repair it."[12] All states have similar laws, such as in Maryland: "[E]very man, for any injury done to him in his person or property, ought to have remedy by the course of the Law of the Land, and ought to have justice and right, freely without sale, fully without any denial, and speedily without delay, according to the Law of the Land."[13]

This right exists around the world, with similar verbiage, in places as diverse as the European Union,[14] countries abiding by Islamic jurisprudence,[15] Israel,[16] and common law regions such as Australia[17] and the United Kingdom.[18] This idea is even bestowed in countries lacking rule of law,[19] whether equally available or not. For example, pre-Taliban Afghanistan,[20] North Korea (although remedies do not exist against an administrative agency in the country),[21] and Egypt[22] all feature a similar right to remedy.

Law, from the origin of the Code of Hammurabi, was made for and is used to ensure social order and disable chaos.[23] It also aims to "care about" the problems of everyone, especially if those individuals are wrong done. Perhaps Hammurabi's code—arguably the first legal code as it was written in 1754 BC in Babylon—was made for such purpose.[24] This idea of ensuring justice from a breaching defendant for all persons

---

[10] LA. CIV. CODE ANN. art. 1756, 1986, 1994, 1809, and 1873.
[11] *See, e.g.*, LA. CIV. CODE ANN. art. 1994 (2011); LA. CIV. CODE ANN. art. 1998 (2011); *see also* Steven J. Burton, *Breach of Contract and the Common Law Duty to Perform in Good Faith*, 94 HARV. L. REV. 369 (1980); Lawrence J. Meyer, *Anticipatory Breach of Contract—Effects of Repudiation*, 8 U. Mia. L. Rev. 68 (1953).
[12] LA. CIV. CODE ANN. art. 2315 (2011).
[13] MD. CONST. art. 19.
[14] Charter of Fundamental Rights of the European Union art. 47, Dec. 14, 2007, 2007 O.J. (C 303) 12 [hereinafter Charter of Rights].
[15] Nabil Saleh, *Remedies for Breach of Contract Under Islamic and Arab Laws*, 4 ARAB L. Q. 269 (1989).
[16] Ernst Livneh, *Criteria of Liability for Breach of Contract*, 2 ISR. L. REV. 67 (1967).
[17] Peter Cane, *Damages in Public Law*, 9 OTAGO L. REV. 489 (1999).
[18] ADVOCS. FOR INT'L DEV., AT A GLANCE GUIDE TO BASIC PRINCIPLES OF ENGLISH CONTRACT LAW (Allen & Overy eds., 2016).
[19] *See* World Bank, *Rule of Law – Country Rankings*, GLOBALECONOMY.COM, https://www.theglobaleconomy.com/rankings/wb_ruleoflaw/ (last visited Dec. 11, 2023).
[20] SAM JACOBSON ET AL., AFG. LEGAL EDUC. PROJECT, AN INTRODUCTION TO THE LAW OF OBLIGATIONS OF AFGHANISTAN (Trevor Kempner et al. eds., 1st ed. 2014).
[21] Jong-Ik Chon, *Basic Rights Under the North Korean Constitution and Related Legal Systems*, 21 J. KOREAN L. 113 (2022).
[22] Ehab Yehia, *Spotlight: Breach of Contract Claims in Egypt*, LEXOLOGY, https://www.lexology.com/library/detail.aspx?g=eb272891-77a0-4ffc-bf98-793c662ed405 (last visited Dec. 11, 2023).
[23] *Hammurabi's Code of Laws*, STUDENTS OF HIST., https://www.studentsofhistory.com/hammurabi-s-code (last visited Dec. 11, 2023).
[24] *Id.*

dates as far back as Hammurabi's code,[25] which focused largely (nearly fifty percent of the code) on contract law, i.e., private law.[26] Roman jurist Ulpian, born in 170, historically dichotomized the idea of private law,[27] although these ideals were already in place in Hammurabi's code. Ulpian investigated not only contract issues, but also more informal protections for single or two-person obligations, such as grain harvesting, gardening, and broker-to-merchant issues.[28]

Since this first example of a legal code provided remedy to all regardless of social status, 3,800 years ago, it can be understood providing such a right is an instinctive idea.[29] If it is, it would mean demanding justice via legal review and the right to remedy is never selfish; the right to remedy reflects law by nature, not a "spoiled" fantasy. However, it is understandable that many citizens under flawed legal systems in countries across the world may fairly think having a right to remedy is a privilege. We will unravel why.

I. COMMON AWARENESS OF THE EXISTENCE OF CONTRACTS AND THE INDEBTED RIGHT TO RESTORATION

The idea of an *owed remedy* in private law, specifically in obligations and contract law, is an antiquated yet globally understood idea in jurisprudence. The concept encapsulated a large chunk of even the first legal code, implying it has always been an idea, an issue, and possibly also an instinctive human right. Due to this, legal remedy is something laypeople assume they are owed in society. Knowing the legal history, can we derive that repudiation and remedy upon a duty being broken in private law is instinctive?

II. INSTINCT OF AN OWED REMEDY IN CONSUMER TRANSACTIONS AND HANDSHAKES

Let's look at consumer transactions and society. Consumers are often aware—even if they don't usually read it—that the "fine print" for products such as a subscription service or cruise ship ticket, etc. is legally binding.[30] Considering the legalese used,[31] consumers believe the "fine print" creates a set of terms they should be subservient to, even though this is not the reality.[32] This subservience suggests that people may view the duties imposed by retail contracts as just and righteous, despite the reality.

---

[25] Patrick J. Kiger, *How the Code of Hammurabi Influenced Modern Legal Systems*, HIST. (Aug. 22, 2023), https://www.history.com/news/hammurabi-code-legal-system-influence.
[26] *Hammurabi's Code*, LUMEN LEARNING, https://courses.lumenlearning.com/suny-hccc-worldcivilization/chapter/hammurabis-code/ (last visited Dec. 11, 2023).
[27] *Private Works Act*, WAYNE J. JABLONOWSKI L., https://wjjlaw.com/lien-bond-claims/private-works-act/ (last visited Dec. 11, 2023).
[28] AVALON PROJECT, THE CODE OF HAMMURABI (L. W. King trans., 2008), https://avalon.law.yale.edu/ancient/hamframe.asp#.
[29] Kiger, *supra* note 25.
[30] Tess Wilkinson-Ryan, *A Psychological Account of Consent to Fine Print*, 99 IOWA L. REV. 1745 (2014).
[31] *See When Is Fine Print a Must-read?*, ARAG LEGAL, https://www.araglegal.com/member/learning-center/topics/budget-and-finance/when-fine-print-must-read (last visited Dec. 11, 2023).
[32] Omri Ben-Shahar, *Fine Print Subservience*, JOTWELL (July 30, 2019), https://contracts.jotwell.com/fine-print-subservience/.

A. Instinctive Owed Remedy 1: Right to Refund, Without Receipt

In contrast, in their consumer roles, do humans feel they deserve justness for their own mistakes, even when they know there is no legal or contractual basis? A common example is asking for a refund despite a lost receipt. Some consumers in retail settings will demand a refund for a purchased item, even if the store has a policy that a receipt must be presented. In making such requests, consumers may consider that a store will offer compassion, either because it understands the likelihood of misplacement or graciously assumes the item was truly purchased and not stolen by the returning consumer. Such a commitment to customer service will encourage a store to do the *right thing*, at least in the customer's mind. Moreover, the consumer may even proclaim these possibilities, or further support their request with assertions such as "I have shopped here for years, you should cut me some slack." A customer suggesting that their legitimacy and loyalty demands retailer flexibility in response to such claims, despite store policy, reveals that the right to remedy might be instinctive.

B. Instinctive Owed Remedy 2: Right to Remedy for Defective Products

Although consumers are likely not well-versed in the realm of products liability law and often do not read "fine print" (nearly half of buyers admit to not reading return policies),[33] they are likely still aware, based on some moral compass, when a product defect is so severe or a product so unusable that they deserve a remedy. For example, if I purchased an item from someone, whether via an informal business deal in cash or by purchasing from a formal retailer, and the product is faulty, I believe I should be entitled to a refund or replacement. In 2022, products liability cases hit an all-time high—nearly 6,000—suggesting this principle is well known despite a lack of consumer legal knowledge.[34]

C. Instinctive Owed Remedy 3: Shaking Hands is a Sign of a Deal

Another example of feeling some internal, instinctive right from an obligee is the idea of shaking hands and thereby creating an agreement with some extra level of security. For example: "We shook hand; thus, you owe me such performance." Shaking hands instinctively signifies the forming of a contract, and a breach of a contract deserves a remedy. That is: "We shook hands, but you didn't deliver; we had a deal so I deserve my obligation to be fulfilled."

Shaking hands has been a way to form a contract since the ninth century.[35] The antiquity of this cultural norm, in conjunction with other previously discussed principles, suggests there are instinctive societal norms about what constitutes justice, regardless of each individual's legal knowledge.

---

[33] Brandon Batchelor, *How Your Return Policy Can Influence New Sales and Long-term Loyalty*, FORBES (June 12, 2020, 8:30 AM), https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2020/06/12/how-your-return-policy-can-influence-new-sales-and-long-term-loyalty/.

[34] *Report: Product Liability Lawsuits Hit Record High in 2022*, KIRKLAND & ELLIS (Sept. 14, 2023), https://www.kirkland.com/news/in-the-news/2023/09/report-product-liability-lawsuits-hit-record-high-in-2022.

[35] *Is a Handshake a Legal Contract*, OBOLOO, https://oboloo.com/blog/is-a-handshake-a-legal-contract/ (last visited Dec. 11, 2023).

Shaking hands is not always a method to properly form a contract;[36] however, it has been cited as an acceptable means of formation in modern case law as recent as September 2023.[37] Thus, shaking hands has legal relevance. Case law has even referenced a "shaking of hands" to signify an agreement,[38] implicating that shaking hands is a bedrock eponym for sealing a contract.

All aforementioned situations—whether demanding a return without a receipt, knowing a person should not be stuck with a broken product, or shaking hands to create a serious obligation—illustrate the instinctive legal trait of an owed remedy. Regardless of store policy or a consumer's policy or legal ignorance, these are all commons features of our social contract. These examples further reinforce the idea that there is a shared understanding of duty: The obligee party receiving the handshake knows they are owed delivery or performance, while the obligor party offering the handshake knows what they do or do not owe, and that the fulfillment of their duty will extinguish the obligation.

III. BELIEF IN A JUST CONSUMER TRANSACTION AND SATISFACTION: THE SIGNIFICANCE OF THE FIRST WRITTEN COMPLAINT

A clay tablet, written in Akkadian merely a few years after Hammurabi's code, was discovered in Ur in present day Iraq.[39] The tablet became a viral phenomenon,[40] and earned the Guinness World Record of the "Oldest Customer Complaint."[41]

The complaint was addressed from a consumer, Nanni, to Ea-nāṣir, a merchant of copper ingot. Nanni sent a servant to purchase the copper from Ea-nāṣir. Ea-nasir delivered the metal both late and of a lower grade than was satisfactory. Ea-nāsir also was supposedly rude to Nanni's servant.

The translated tablet written to Ea-nāṣir reads:

> *"Tell Ea-nāsir Nanni sends the following message: When you came, you said to me as follows: 'I will give Gimil-Sin (when he comes) fine quality copper ingots.' You left then but you did not do what you promised me. You put ingots which were not good before my messenger (Sit-Sin) and said: 'If you want to take them, take them; if you do not want to take them, go away!' What do you take me for, that you treat somebody like me with such contempt? I have sent as messengers gentlemen like ourselves to collect the bag with my money (deposited with you) but you have treated me with contempt by sending them*

---

[36] *Id.*
[37] Vukadinovich v. Posner, No. 2:22-CV-118-TLS-JPK, 2023 WL 6211835, at *1 (N.D. Ind. Sept. 25, 2023) ("The Plaintiff agreed to the amendments, and the Plaintiff and the Defendant shook hands.").
[38] Dist. 4, Commc'ns Workers of Am., AFL-CIO v. NLRB, 59 F.4th 1302, 1312 (D.C. Cir. 2023) (finding a "meeting of the minds" occurs "where an employer's 'remarks . . . were the email equivalent of *shaking hands* on the deal at the end of a face to face meeting'") (emphasis added).
[39] *Oldest Written Customer Complaint*, GUINNESS WORLD RECORDS, https://www.guinnessworldrecords.com/world-records/537889-oldest-written-customer-complaint (last visited Dec. 11, 2023).
[40] Christina Zhao, *3,800-year-old Tablet with World's Oldest Customer Complaint Goes Viral: 'What Do You Take Me For?'*, NEWSWEEK (Aug. 24, 2018, 8:56 AM), https://www.newsweek.com/3800-year-old-tablet-worlds-oldest-customer-complaint-goes-viral-who-do-you-1088904.
[41] *Oldest Written Customer Complaint*, *supra* note 39.

> *back to me empty-handed several times, and that through enemy territory. Is there anyone among the merchants who trade with Telmun who has treated me in this way? You alone treat my messenger with contempt! On account of that one (trifling) mina of silver which I owe(?) you, you feel free to speak in such a way, while I have given to the palace on your behalf 1,080 pounds of copper, and umi-abum has likewise given 1,080 pounds of copper, apart from what we both have had written on a sealed tablet to be kept in the temple of Samas. How have you treated me for that copper? You have withheld my money bag from me in enemy territory; it is now up to you to restore (my money) to me in full. Take cognizance that (from now on) I will not accept here any copper from you that is not of fine quality. I shall (from now on) select and take the ingots individually in my own yard, and I shall exercise against you my right of rejection because you have treated me with contempt.*"[42]

"The Complaint Tablet to Ea-nāṣir," etched in 1750 BC, serves as a remarkable ode to the timeless foundations of contract law, but also to the generally assumed rights of an obligee which span millennia. Despite the vast temporal expanse, this ancient exchange remarkably resonates with the core elements of contemporary contract law. It encapsulates the essence of contract initiation, violation, and dispute resolution—a tapestry that has endured through the annals of history. The timeline and remarkable relatability of this complaint suggest equity is a legal principle, but really—an instinct at heart. This suggests that such complaints based upon the expectations of people in society to receive their fair end of the bargain in contractual relationships, a quality product, a chance for a fair remedy when they are owed one, as well as a standard of business against which the obligor is not immune, are not nuanced.

In the 3,777 years since the society of Old Babylonian thrived, law and society have evolved. Yet we now find ourselves in the United States still adhering to similar legal principles. This emphasizes our enduring expectations for equitable consumer experiences and the belief that every consumer should be treated fairly and provided with equitable contract terms, ensuring they don't end up with subpar products or fall victim to deception.

Then and now, the emotional experiences in retail and contractual situations remain straightforward. The notion that your concerns should be taken seriously and that those responsible for any wrongdoing should be held accountable is clear and straightforward. However, this notion has been emotionally undermined, and therefore made societies not consider the right to remedy an entitlement, in the governmental situations present in countries outside the United States. Such situations contrast against the belief in the prevalence of impartial justice and the commitment of the judiciary and government to fairness, addressing transgressions, and adhering to established norms, including expectations and the rule of law. This contrast also encompasses the government's obligation to respond to allegations of wrongdoing and fulfill its responsibilities promptly, such as rectifying delays, addressing corruption among officers, rectifying negligence within the military, and addressing bribery requests by judges.

To put it succinctly, if a situation like those brought under the Federal Tort Claims Act were to occur in the United States versus a country lacking a strong rule of

---

[42] *Id.*

law (in this example, referred to as Country X), the transparency and expectation of justice in the United States would significantly impact an American plaintiff's outlook whereas the lack of such principles would negatively zffect a plaintiff in Country X.

An American in this scenario would have much higher expectations and confidence in the delivery of justice. They would be far less likely to tolerate misconduct from the judiciary and would insist on and expect equal measures of judicial remedies for issues such as delays, bribery, and injustice. ("I have been wrong done, therefore, because it is wrong, and justice is fair, then I will receive justice:" This is the peak idea that deteriorates under regimens facing rule of law issues.)

IV. WHAT LEADS TO THE DIMINISHING TRUST AND CONVICTION THAT JUSTICE WILL BE SERVED AND SHOULD BE EXPECTED?

Since the entitled belief that obligees are owed benefits, remedy, and equity, at least from private parties in private law, particularly in contract law, is not a nuanced nor purely western idea—in fact, obligees' rights to performance and remedy are codified across the globe—why is such justice unavailable in places such as Sudan,[43] Ghana,[44] Somalia,[45] Namibia,[46] Albania,[47] Indonesia,[48] Morocco,[49] Malawi,[50] and Afghanistan,[51] to name only a small handful? This can be seen through protests, proposals for restorative legislation, reports, surveys, and other data.

Contract law is wonderful as it protects the interests of small claims and thus provides justice among social classes. However, in some countries like South Sudan,[52] the judiciary has made such protections a challenge for poorer people. In South Sudan, courts are highly criticized for issues such as bribery, favoritism, and long delays, which particularly affect the poor and aggravate conflicts. Government courts face the brunt of these criticisms as due process requirements are often viewed as breeding corruption and escalating disputes. Obstacles to justice include government and military interference, police misconduct, weak enforcement, and a perceived erosion of traditional authority. Despite these challenges, litigants in South Sudan pragmatically choose between restorative and adversarial dispute resolution methods, depending on

---

[43] *See Sudan Appoints First Female Judiciary Head to Fight Corruption*, REUTERS (Oct. 10, 2019, 5:58 AM), https://www.reuters.com/article/us-sudan-politics/sudan-appoints-first-female-judiciary-head-to-fight-corruption-idUSKBN1WP2DM/; https://www.ganintegrity.com/country-profiles/sudan/.

[44] *See* Franck Kuwonu, *Judiciary: Fighting Graft Needs Muscles*, AFR. RENEWAL (Aug.–Nov. 2016), https://www.un.org/africarenewal/magazine/august-2016/judiciary-fighting-graft-needs-muscles.

[45] *See Freedom in the World 2022: Somalia*, FREEDOM HOUSE, https://freedomhouse.org/country/somalia/freedom-world/2022 (last visited Dec. 11, 2023).

[46] *See Namibia Risk Report*, GAN INTERGRITY (Nov. 4, 2020), https://www.ganintegrity.com/country-profiles/namibia/.

[47] *See* Benet Koleka, *Scuffles, Flares as Albania Picks Interim Prosecutor*, REUTERS (Dec. 18, 2017, 5:15 AM), https://www.reuters.com/article/us-albania-prosecutor-protests/scuffles-flares-as-albania-picks-interim-prosecutor-idUSKBN1EC1XH.

[48] *See* MAIRA MARTINI, TRANSPARENCY INT'L, CAUSE OF CORRUPTION IN INDONESIA (2012).

[49] *See Morocco Risk Report*, GAN INTERGRITY (Nov. 4, 2020), https://www.ganintegrity.com/country-profiles/morocco/.

[50] *See Malawi Anti-bribery Protests Draw Thousands*, VOA NEWS (Jan. 16, 2020, 6:58 PM), https://www.voanews.com/a/africa_malawi-anti-bribery-protests-draw-thousands/6182719.html.

[51] *See* MARIE CHÊNE, TRANSPARENCY INT'L, TACKLING JUDICIAL CORRUPTION IN AFGHANISTAN (2007).

[52] CHERRY LEONARDI ET AL., PEACEWORKS, LOCAL JUSTICE IN SOUTHERN SUDAN (2010).

the specific circumstances and social dynamics at play.[53] South Sudan is not an uncommon situation. Comparable or identical systems can ferment and transpire anywhere where bribery is required for speedy and righteous justice. It is important to note that many countries who experience issues of bribery in their court systems are also often impoverished.[54] Poor people cannot compete in systems where justice is swayed by financial bribery or influence.[55] This is one example of how, although a belief in equal scales in private law may be instinctual, one's experience within a particular government's judicial system may cause, at minimum, a diminishing belief in the judiciary. In particular, those living under corrupt judicial systems may lack faith that the court itself will help render reflective judgments and do the *right thing.*

Weak rule of law and corruption in countries affect access to remedy.[56] It is said that "[i]f money and influence are the basis of justice, the poor cannot compete."[57] These two issues work together against poorer people. Further, even if law exists on paper in a weak rule of law country, it is likely not practiced entirely justly. For example, duress is prohibited in the constitution of Iran,[58] yet there have been instances where the government itself has coerced the accused.[59] Further, the constitutions of Iran, Somalia, and Albania all vow to uphold justice,[60] yet each nation has had cited issues, in practice, of lacking due process.[61]

Although many countries have anti-corruption laws, such as Somalia, Zimbabwe, and Morocco,[62] they still have cultures of bribery within the judicial

---

[53] *Id.* at 52.

[54] *Judicial Corruption Fuels Impunity, Corrodes Rule of Law, Says New Transparency International Report*, TRANSPARENCY INT'L (May 23, 2007), https://www.transparency.org/en/press/20070523-judicial-corruption-fuels-impunity-corrodes-rule-of-law-says-repor; *see also Rule of Law – Country Rankings*, GLOBALECONOMY.COM, https://www.theglobaleconomy.com/rankings/wb_ruleoflaw/ (last visited Dec. 11, 2023); *Poorest Countries in the World 2023*, WORLD POPULATION REV., https://worldpopulationreview.com/country-rankings/poorest-countries-in-the-world (last visited Dec. 11, 2023).

[55] *Judicial Corruption Fuels Impunity*, *supra* note 54.

[56] GWYNNE L. SKINNER ET AL., TRANSNATIONAL CORPORATIONS AND HUMAN RIGHTS: OVERCOMING BARRIERS TO JUDICIAL REMEDY (2020).

[57] *Judicial Corruption Fuels Impunity*, *supra* note 54.

[58] Nicolas Garon, *Veiling Laws Throughout Iranian History: The Relationship to Religion, Before and During Islamic Law*, 38 CONN. J. INT'L L. 50, 60 (2023).

[59] Rosie Swash, *Arrests and TV Confessions as Iran Cracks Down on Women's 'Improper' Clothing*, GUARDIAN (Aug. 23, 2022, 1:30 PM), https://www.theguardian.com/global-development/2022/aug/23/arrests-and-tv-confessions-as-iran-cracks-down-on-women-improper-clothing-hijab; *Respect Lives, Voices of Iranians and Listen to Grievances, Pleads UN Human Rights Chief*, U.N. OFF. HIGH COMM'R HUM. RTS. (Jan. 10, 2023), https://www.ohchr.org/en/press-releases/2023/01/respect-lives-voices-iranians-and-listen-grievances-pleads-un-human-rights; *see also* Garon, *supra* note 58.

[60] Islahat Va Taqyyrati Va Tamimah Qanuni Assassi [Amendment to the Constitution] 1368 [1989] (Iran); ALB. CONST. 1998 (amended 2016); SOM. CONST. 2012.

[61] *See, e.g.*, *Respect Lives, Voices of Iranians*, *supra* note 59; *Freedom in the World 2022: Albania*, FREEDOM HOUSE, https://freedomhouse.org/country/albania/freedom-world/2022 (last visited Dec. 11, 2023); *Somalia: Events of 2021*, HUM. RTS. WATCH, https://www.hrw.org/world-report/2022/country-chapters/somalia (last visited Dec. 11, 2023).

[62] MARIE CHÊNE, TRANSPARENCY INT'L, OVERVIEW OF CORRUPTION AND ANTI-CORRUPTION IN SOMALIA (2012); *Zimbabwe – Global Bribery Offenses Guide*, DLA PIPER (Jan. 11, 2022), https://www.dlapiper.com/en/insights/publications/2019/09/global-bribery-offenses-guide/zimbabwe; *Anti-corruption in Morocco*, BAKER MCKENZIE: GLOBAL COMPLIANCE NEWS BLOG https://www.globalcompliancenews.com/anti-corruption/anti-corruption-in-morocco/ (last visited Dec. 11, 2023).

arena.[63] Bribery is common: In 2007, a paper found that by bribing judges directly, a party could either delay or accelerate their case in thirty-two countries.[64] Sudanese courts feature "bribery, favoritism, and excessive delays, which significantly disadvantage the poor."[65] Similarly, stalling bribes have been uncovered in India.[66] In the legal system of Zimbabwe, bribery can stall cases for so long it makes "plaintiffs . . . frustrated enough to withdraw their case" and leads "the media and public [to] no longer [be] interested in the outcome."[67]

Such deterioration of interest—both from plaintiffs and society at large—as caused by human interference with a fair and unbiased judiciary ultimately results in a loss of faith in justice; in such a corrupt system justice is not accessible and poorer people's voices cannot be heard. The United Nations Office on Drugs and Crime found "[c]orruption undermines the core of the administration of justice, generating a substantial obstacle to the right to an impartial trial, and severely undermining the population's trust in the judiciary." It can also perpetuate a viewpoint that government bodies lack accountability and can be reckless.

Bribes were demanded in Iran from judges to release protestors.[68] Bribery also affects justice in Ghanaian,[69] Mozambican,[70] and Burmese[71] courts, and direct bribes to judges and/or magistrates have happened in Afghanistan, [72] Somalia, [73] and Bangladesh,[74] or to the court generally in Malawi, South Africa, and Namibia[75]. Over half of those who received a judicial service in Bangladesh had to pay a bribe.[76]

Transparency International Found that "in more than twenty-five countries, at least one in ten households had to pay a bribe to get access to justice. In a further twenty

---

[63] Maxamed Mubarak, *Judicial Corruption in Somalia*, MARQAATI (Nov. 6, 2014), https://marqaati.org/en/2014/11/judicial-corruption-in-somalia/; Tracy Mutowekuziva, *Delivering Justice in Zimbabwe's Courts*, TRANSPARENCY INT'L: BLOG (June 10, 2020), https://www.transparency.org/en/blog/delivering-justice-in-zimbabwes-courts; *Judicial Corruption Fuels Impunity*, *supra* note 54.

[64] *Judicial Corruption Fuels Impunity*, *supra* note 54.

[65] LEONARDI, *supra* note 52, at

[66] Vani S. Kulkarni et al., *India's Judiciary and the Slackening Cog of Trust*, HINDU (May 9, 2022, 12:06 AM), https://www.thehindu.com/opinion/op-ed/indias-judiciary-and-the-slackening-cog-of-trust/article65394817.ece#.

[67] Mutowekuziva, *supra* note 63.

[68] Jubin Katiraie, *Iran's Judges Demand Bribes for Protester's Release*, IRAN FOCUS (Jan. 31, 2020), https://iranfocus.com/protests/34247-iran-protest-judge-bribes-20200131/.

[69] MORGAN BRIGHT GORDON, BRIBERY AND CORRUPTION IN PUBLIC SERVICE DELIVERY: EXPERIENCE FROM THE GHANA JUDICIAL SERVICE (2017).

[70] *Mozambique Risk Report*, GAN INTERGRITY (Nov. 5, 2020), https://www.ganintegrity.com/country-profiles/mozambique/.

[71] Zue Zue, *Burma's Judicial System Deeply Corrupt, Parliament Told*, IRRAWADDY (Dec. 9, 2015), https://www.irrawaddy.com/news/burma/102553.html.

[72] U.N. OFF. DRUGS & CRIME, CORRUPTION IN AFGHANISTAN: BRIBERY AS REPORTED BY THE VICTIMS (2010).

[73] Mubarak, *supra* note 63.

[74] *Bangladesh Risk Report*, GAN INTERGRITY (Nov. 4, 2020), https://www.ganintegrity.com/country-profiles/bangladesh/.

[75] Carmel Rickard, *State of the Judiciary: New Report on Malawi, Namibia, South Africa*, AFRICAN LII (May 6, 2022), https://africanlii.org/articles/2022-05-06/carmel-rickard/state-of-the-judiciary-new-report-on-malawi-namibia-south-africa.

[76] FARZANA NAWAZ, TRANSPARENCY INT'L, OVERVIEW OF CORRUPTION WITHIN THE JUSTICE SECTOR AND LAW ENFORCEMENT AGENCIES IN BANGLADESH (2012).

countries, more than three in ten households reported that bribery was involved in securing access to justice or a 'fair' outcome in court. In Albania, Greece, Indonesia, Mexico, Moldova, Morocco, Peru, Taiwan[,] and Venezuela, the figure was even higher."[77]

Whether corrupt due to economic standing, political bias and influence, or merely a judge's interest in personal gain, such cultures of bribery allow money to influence judicial services, case outomces, and ultimately justice\. These bribes and biases immediately disadvantage poor and ordinary people, favoring those with money or power. This delegitimatizes the judiciary, which lacks impartiality under such a corrupt culture. Thereby, the citizenry of such places lose faith in the courts, and also experience a bitter feeling of injustice.

V. NON-STATE JUDICIARIES

All of the aforementioned countries are recognized states. The question is then, what about non-state legal systems? The United Nations affords protections to tribal minorities within countries.[78] Moreover, many countries themselves give tribal groups within their borders the right to legal protection.[79] (However, other countries have laws which allowtribes to have their own courts and sometimes parliaments.)[80]

In Pakistan, the former "Federally Administered Tribal Area"—which now has been merged into Khyber Pakhtunkhwa, a region known for lawlessness and harboring terrorists[81]—would, from an outward eye, seem to be the wild west of legal rule.

In areas like this in Pakistan and Afghanistan, tribal disputes are sometimes heard from a meeting of village elders known as a jirga.[82] The jirga, which bases its

---

[77] *Judicial Corruption Fuels Impunity*, *supra* note 54.

[78] G.A. Res. 61/295, Declaration on the Rights of Indigenous Peoples (Sept. 13, 2007).

[79] *See, e.g.*, ABBI BUXTON & EMMA WILSON, FPIC AND THE EXTRACTIVE INDUSTRIES: A GUIDE TO APPLYING THE SPIRIT OF FREE, PRIOR AND INFORMED CONSENT IN INDUSTRIAL PROJECTS, (2013).

[80] *See, e.g.*, *Tribal Courts*, U.S. BUREAU JUST. STAT. https://bjs.ojp.gov/topics/tribal-crime-and-justice/tribal-courts (last visited Dec. 11, 2023); The Indigenous Peoples' Rights Act, Rep. Act No. 8371, (Oct. 29, 1997) (Phil.); Apoorv Kurup, *Tribal Law in India: How Decentralized Administration Is Extinguishing Tribal Rights and Why Autonomous Tribal Governments Are Better*, 7 INDIGENOUS L.J. 87, (2008); Michele Langevine Leiby, *In Pakistan, a Legal System Under Scrutiny*, WASH. POST (May 29, 2012, 9:46 AM), https://www.washingtonpost.com/world/asia_pacific/in-pakistan-a-legal-system-under-scrutiny/2012/05/29/gJQAmJTqyU_story.html; *South Africa: Legal Resources: Customary Law & Indigenous Peoples*, BODLEIAN LIBRS., https://libguides.bodleian.ox.ac.uk/law-s_africa/indigenous (last visited Dec. 11, 2023); *Using Tribal Court (Kgotla) for Consultation and Decision-making*, REPUBLIC BOTSWANA, http://www.gov.bw/public-safety/using-tribal-court-kgotla-consultation-and-decision-making (last visited Dec. 11, 2023); U.N. Permanent F. Indigenous Issues, Response dated Feb. 17, 2016 from the Sami Parliament of Norway addressed to the U.N. Permanent Forum on Indigenous Issues, U.N. DOC. 16/712–4 (Feb. 17, 2016); *Background: The State and the Sami Parliament*, SÁMEDIGGI, https://www.sametinget.se/9688 (last visited Dec. 11, 2023); *Quienes Somos*, CONAMAQ, https://www.conamaq.org/quienes-somos/ (last visited Dec. 11, 2023); *¿Qué es el CNI?*, CONGRESO NACIONAL INDÍGENA, https://www.congresonacionalindigena.org/que-es-el-cni/ (last visited Dec. 11, 2023).

[81] Zahid Hussain, *Pakistan's Most Dangerous Place*, WILSON Q., https://www.wilsonquarterly.com/quarterly/_/pakistans-most-dangerous-place (last visited Dec. 11, 2023).

[82] ELI SUGARMAN ET AL., AFG. LEGAL EDUC. PROJECT, AN INTRODUCTION TO THE COMMERCIAL LAW OF AFGHANISTAN (Daniel Lewis et al. eds., 2d ed. 2011); AMNESTY INT'L, PAKISTAN: THE TRIBAL JUSTICE SYSTEM, (2002).

decisions off tribal custom, have been ruled illegal in Pakistan,[83] and have also been called a transgression of human rights.[84] Yet, jirgas often still hear private law matters and offer restorative justice.[85]

Tribal courts in Botswana,[86] Jordan,[87] and Afghanistan[88] also handle private, contract, and commercial law issues alongside the state judiciaries. In fact, in places like Afghanistan, people rely more on tribal jirgas than the state-sanctioned judicial authority due to challenges within the formal legal system.[89]

This furthers the claim that even in such jurisdictions, private law, including contract or any other person-to-person matter, is of such importance, it must be heard. Thereby, it continues to exist within legal practice even if the formal judicial system is lacking. Arguably, the only examined country which blockades citizens from exercising any sort of legal authority in aim of a remedy is North Korea, where actions cannot be made against the state.[90]

CONCLUSION

From Hammurabi's intent to the present day, from tribal to authoritarian law, from democratic countries to places lacking freedom, the protection of the interests of individuals in private law are common, span through time, and are seen in all legal systems. The earliest legal principles pertaining to private law were established to ensure the rights of ordinary people in transactions were upheld transparently, equitably, and without influence or requirement of money. As society progresses, common activities and their issues birth reflective need for contract law regulations—from grain issues in Old Babylonia, to chip packaging patents,[91] to edge contracts today.[92] It can be inferred that throughout history, a fundamental human inclination toward fairness within private law, especially contract law, has existed. This is accompanied by an expectation of the rule of law, the safeguarding of the rights of parties to an obligation, and a defense against unjust treatment. Such principles and the overlying belief in access to remedy are more expected to be upheld in places with a proper rule of law and pure judiciary.

---

[83] Momina Khurshid, *Jirga System in Pakistan: A Transgression of Human Rights*, RSCH. SOC'Y INT'L L. (Apr. 11, 2022), https://rsilpak.org/2022/jirga-system-in-pakistan-a-transgression-of-human-rights/.
[84] AMNESTY INT'L, *supra* note 82.
[85] Zia Akhtar, ADR (Grand Jirga), Truth, Justice and Reconciliation Commission and Peace on the Pakistan-Afghanistan Frontier (Ph.D. dissertation, Sussex University) (on file with Human Rights Law Review, Queen Mary University of London); NAVEED AHMAD SHINWARI, UNDERSTANDING JIRGA: LEGALITY AND LEGITIMACY IN PAKISTAN'S FEDERALLY ADMINISTERED TRIBAL AREAS, (2011).
[86] U.S. AGENCY INT'L DEV., USAID COUNTRY PROFILE – PROPERTY RIGHTS AND RESOURCE GOVERNANCE: BOTSWANA; Piwane Constance Moumakwa, The Botswana Kgotla System: A Mechanism for Traditional Conflict Resolution in Modern Botswana. Case Study of the Kanye Kgotla, (Autumn 2010) (M.A. Thesis, University of Tromsø) (on file with Artic University of Norway).
[87] Ann Furr & Muwafaq Al-Serhan, *Tribal Customary Law in Jordan*, S.C. J. INT'L L. & BUS., Spring 2008.
[88] SUGARMAN, *supra* note 82.
[89] *Id.* at 88.
[90] Chon, *supra* note 21.
[91] *See generally* Max A. Cherney, *TSMC Leads in Advanced Chip Packaging Wars, LexisNexis Patent Data Says*, REUTERS (Aug. 1, 2023, 7:09 AM), https://www.reuters.com/technology/tsmc-leads-advanced-chip-packaging-wars-lexisnexis-patent-data-says-2023-08-01/.
[92] *See generally* Spencer Williams, *Edge Contracts*, 25 U. PENN. J. BUS. L. 839 (2023).

These principles, upon which our legal systems are founded, and which aim to foster a better society, are not novel concepts. They are not limited to first world nations but are historical and seemingly instinctive regardless of society or date.

However, although this feeling of a right to remedy may be instinctive, it is sometimes eroded, or totally deflated, by corruption or political influence within the judiciary itself. Corruption and bribery are regrettably prevalent in regions where the rule of law is lacking, despite such countries often having a legal code which purports to uphold due process. Consequently, this undermines the prospects of a fair trial for individuals from lower socioeconomic backgrounds and may discourage the pursuit of small claims, including disputes arising from single-party contracts.

The desire for a remedy in any government at any time period is never a spoiled idea. Oversight and strong-willed people demanding concessions and change can beat a stagnant and money hungry judiciary. Since the instinct for justice is forever engrained, limits on it shall not sustain.

THE LEGAL FRAMEWORK FOR CROSS-BORDER DATA TRANSFER BETWEEN MAINLAND CHINA AND HKSAR

Junxuan Wu*

Abstract: Cross-border data transfer is a hard issue in today's world of "digital nationalism". In this post-Snowden world, data-localization has become the norm. China has adopted data localization rules in various laws, from Internet Security Law to Data Security Law. China's constitutional structure of "one country, two systems" presents a unique question to data localization: should cross-border data transfer between the Mainland and SARs (Special Administrative Regions) be constrained by data-localization rules? Since both basic laws for Hong Kong and Macau define these two SARs as "free trade" zones and "separate customs" territories, once data from the Mainland are transferred to the SARs, there would be no existing laws to hinder their further flow to the globe. Furthermore, the SARs have their laws protecting data rights and regulating data use, which are quite different from the national laws. These unique features render cross-border data transfer *within* China a challenging and interesting topic. This article takes the challenge by focusing on the legal framework for data transfer between Mainland China and Hong Kong. It delineates the relevant legal rules in China and its HKSAR, points out the obstacles and difficulties, and suggests reforms.

Keywords: Cross-Border Data Transfer; Digital Sovereignty; Data Localization; National Security; Hong Kong SAR; Mainland China

* KoGuan Law School, Shanghai Jiao Tong University, China.

Table of Contents

INTRODUCTION

It is often believed that the Internet is borderless and data flow is free. However, this belief is untrue from both technical and legal perspectives. Technically, data flow is controlled by border gateway protocol and firewalls. Legally, it is regulated by data transfer rules. Actually, tightening border control on Internet is the general trend in the whole world except in the United States, because the U.S. Internet hegemony is the reason for all other countries to defend their digital border by making data localization rules. Anupam Chander called this trend "data nationalism" and made a general description: "The era of a global Internet may be passing. Governments across the world are putting up barriers to the free flow of information across borders. Driven by concerns over privacy, security, surveillance, and law enforcement, governments are erecting borders in cyberspace, breaking apart the World Wide Web. The first generation of Internet border controls sought to keep information out of a country-from Nazi paraphernalia to copyright infringing material.' The new generation of Internet border controls seeks not to keep information out but rather to keep data in. Where the first generation was relatively narrow in the information excluded, the new generation seeks to keep all data about individuals within a country."[1333]

On the other hand, data flow is gradually surpassing traditional cross-border trade of goods and investment, becoming a new driving force for global economic growth. Today, with the vigorous development of the digital economy, cross-border data activities are becoming more and more frequent, and the demand for data outbound transfer by data processors is growing rapidly. Major countries and regions in the world have made various bilateral and multilateral legal arrangements to facilitate cross-border data transfer and used "adequacy" standard to make sure that data trading partners have laws adequate to protect personal information rights and interests.[1334] Legal tools such as TIA (Transfer Impact Assessment) have been developed to address the issue of balancing data trade and data security.

As a major digital economy country, China especially needs to promote cross-border data flow. On the other hand, given the current situation of U.S.-China relationship, China also needs to make sure that its outbound data flow should not undermine its national security and citizen's personal information rights. Since the listing of "Didi Chuxing" in the United States was urgently suspended,[1335] it is no longer feasible for companies to list in the United States due to national data security considerations; on the other hand, due to the mainland's favorable attitude towards listing in Hong Kong, attracted Didi and other data-related companies to turn their attention to Hong Kong, and consider listing in Hong Kong as the main way to obtain global investment. A large number of companies are listed in Hong Kong, and huge amounts of data and information flow seamlessly between the two regions every day, which brings about the legal issues of cross-border data flow between mainland China and Hong Kong. The Hong Kong Special Administrative Region and the Mainland are in different legal jurisdictions, and the flow of data between the two places constitutes cross-border flow. However, current academic research mainly focuses on the study of

---

[1333] Chander, A. and Le, U. P. (2015), Data Nationalism. *Emory Law Journal*, 64(3), 677-740.
[1334] Taylor, Mistale (2023). *Transatlantic Jurisdictional Conflicts in Data Protection Law: Fundamental Rights, Privacy and Extraterritoriality*. Cambridge University Press. 193-195.
[1335] Zhang, Angela Huyue (2024). *High Wire: How China Regulates Big Tech and Governs Its Economy*. Oxford University Press. 57.

cross-border data flow between different countries. Regarding the flow of data between Mainland China and Hong Kong HKSAR, two jurisdictions within one country, there is almost no academic research. This paper intends to fill in the gap.

It starts from a general description of the laws and regulations regarding data of the two places themselves, sorts out and analyzes their regulations and restrictions on data exports, including China's legal control of mainland data retrieval, then compares their regulatory thresholds, regulatory intensity, regulatory purposes, etc.. Next, it identifies the practical problems in cross-border flows from mainland China to the Hong Kong Special Administrative Region. Finally, it proposes certain reforms to address these problems.

I. Mainland China’s Regulations and Restrictions on Cross-border Data Transfer

A. Definition of Data Export

1. Definition of Data

To clarify what "data transfer" is, we first need to clarify what data is. Regarding the definition of data, academic circles have given many opinions from different angles, such as “an information carrier designed to record the subjective reflection of the subject of knowledge on the object of knowledge” and “information is the expression of knowledge and the reaction of the human brain to data. Data is “the embodiment of information” “massive, high-growth and diversified information assets” and so on. EU Database Directive defines data as “any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audio-visual recording.” However, these definitions are all derived from papers published before 2021. Since there has been a clear definition of data in specific laws, it is proper for this article to adopt legal definitions. According to Article 3 of the "Data Security Law of the People's Republic of China" (hereinafter referred to as DSA), data is "any record of information in electronic or other forms." This legal definition gives data a very broad scope, and any record of information electronically or otherwise can be considered data. It is worth noting that data can be divided into many types from different perspectives. For example, based on the subject or purpose, it can be divided into "personal data, business data, technical data, and organizational (public) data". Different types of data, when transferring cross border, may bring about different kinds of legal issues, from privacy, trade secrets to national security. Article 21 of DSA provides that "the state establishes a data classification and hierarchical protection system, and implements classified and hierarchical protection of data according to the importance of data in economic and social development." It suggests that different data, according to their different social values, shall enjoy different levels of protection and regulation. To implement the classification-based and hierarchical regulatory system, detailed regulations have been made.

2. Two Definitions of Cross-Border Data Transfer

Among several existing studies on cross-border data flows, there are two main

mainstream definitions of data export. The first definition is taken from an article published by the United Nations Center on Transnational Corporations in 1984, which defines "Cross-border data flow" as the situation in which "electronic information records generated in one country are read, stored, used or processed by private entities or public authorities in other countries".[1336] This definition emphasized the "transnational" nature of cross-border data transfer.[1337] However, in China's "One Country, Two Systems" constitutional order, there are borders within one country. Hong Kong and Macau, as China's Special Administrative Regions established in accordance with Article 31 of the Constitution, maintain their unique legal systems. Therefore, data transfer between the Mainland and the two SARs should be considered as "cross-border" data transfer. Therefore, "cross-border data transfer" should be redefined as "data generated in one jurisdiction are processed by persons and entities in other jurisdiction(s)".

This definition is not clearly expressed in relevant laws and regulations in China. For example, the Guidelines for Data Export Security Review defines data export as "a one-time or continuous activity in which a network operator provides personal information and important data collected and generated during its operations within the territory of the People's Republic of China to institutions, organizations or individuals outside the country through the Internet or other means, by directly providing or conducting business, providing services or products, etc." Here, "export" means leaving "the territory of the People's Republic of China". A literal reading of this definition may lead to the conclusion that data transfer between Mainland China and HKSAR doesn't constitute data export, because data is still within the territory of China. This kind of confusion is quite common among foreign observers of Chinese law. For example, while discussing the four conditions imposed by the Personal Information Protection Law (PIPL) on cross-border transfer of personal information, Graham Greenleaf cautioned: "It is possible but uncertain that this prohibition might also include Hong Kong."[1338] For any Chinese lawyer, this kind of uncertainty doesn't exist. Legal rules are located in a system of laws under a Constitution. Systematic interpretation of specific rules solves such uncertainty. Article 31 of the Constitution authorizes the National People's Congress (NPC) to establish Special Administrative Regions and their applicable laws. The Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China (hereinafter referred to as the "Basic Law") is a law passed by the NPC and applied to Hong Kong. The Basic Law makes sure that Hong Kong not only enjoys a high degree of autonomy, but also implements laws that are different from those in mainland China. Within such a unique constitutional framework called "One Country, Two Systems", a range of laws, from immigration and border control laws to trade-related laws, established borders between the Mainland and Hong Kong. For example, Article 89 of the Border Exit and Entry Administration Law (《出入境管理法》) defines border exit (出境) as traveling from the mainland of China to other countries or regions, including traveling from the mainland of China to the Hong Kong Special Administrative Region. Obviously, if the traveling of natural persons from the mainland to Hong Kong is crossing the border, traveling of data from mainland

[1336] UNCTC. Transnational Corporations and Transborder Data Flows. The United Nations, 1984.
[1337] Arner, Douglas W., Castellano, Giuliano G., & Selga, Eriks. The Transnational Data Governance Problem. *Berkeley Technology Law Journal*, 2021, 37(2): 623-699.
[1338] Greenleaf, Graham. Personal Data Localization and Sovereignty along Asia's New Silk Roads. In Chander, Anupam, & Sun, Haochen (eds.), *Data Sovereignty: From the Digital Silk Road to the Return of the State*. Oxford University Press, 2023. 301.

to Hong Kong should also be considered as cross-border transfer. Therefore, the conditions imposed by Article 38 of PIPL on cross-border data transfer clearly apply to mainland China-Hong Kong data transfer, including (1) Passing a security assessment organized by Cyberspace Administration of China (CAC) following the provisions of Article 40 of PIPL; (2) Obtaining personal information protection certification through a professional institution; (3) Entering into a standard contract (formulated by CAC) with the overseas recipient, stipulating the rights and obligations of both parties; (4) Other conditions prescribed by laws, administrative regulations or CAC rules.

B. Development of Legislations and Regulations on Cross-border Data Transfer in China

1. Early Data Export Legislation and Characteristics

Article 59 of the National Security Law is the earliest law regarding the security supervision of data exports, but it does not specifically mention data export. This article believes that the characteristics of mainland China's early cross-border data transfer laws can be characterized as low-level, fragmented, narrow coverage, weak operability, and low flow permission. First, the level of the rules is low. There is no national laws to stipulate rules on data transfer. The most high-level rules before the "13th Five-Year Plan" are only administrative regulations in nature, followed by "notices" in the rank of “other normative documents", national standards, and even non-standard documents. It has a mandatory effect and the legislative level is generally low. Second, fragmentation is caused by low rank, because the unity of law is achieved in China by a hierarchical structure stipulated in the Law on Legislation with the Constitution on the top. Fragmentation is reflected in the fact that the regulations on the export of different types of data are scattered in different notices, regulations, and technical documents. For example, the outflow of personal financial information is regulated by the "Notice of the People's Bank of China on Banking Financial Institutions Good Practices in Protecting Personal Financial Information", while the data held by credit reporting agencies is regulated by the "Regulations on the Administration of the Credit Reporting Industry". Because a single document cannot regulate all data exports, there are situations where different documents regulate different special fields. Moreover, these documents are concentrated in the special fields of financial and transportation credit reporting and are narrow in scope and not comprehensive enough. Third, weak operability and low mobility permissions are reflected in the fact that most of the regulations are broad rough, sometimes simply stipulate that data "should be within the country" and "should not be provided overseas". There is a very obvious tendency for data to be stored locally, even if there are exceptions. In situations such as "unless otherwise provided" and "statutory permission", these exceptions have not been specifically refined.

2. The Gradual Formation and Preliminary Improvement of Data Export Legislations

In 2016, the State Council issued the "Thirteenth Five-Year Plan for National Informatization", which clearly stated the strategic requirement of "establishing a security supervision system for cross-border data flows." Since then, legislative supervision of data exports has been gradually and systematically established and

improved. In November 2016, the "Cybersecurity Law of the People's Republic of China" (hereinafter referred to as the "Cybersecurity Law") was promulgated, and the provisions of Article 37 reflect the requirements for data localization. Data can only be exported abroad if it is truly necessary to provide it overseas and if it passes the security assessment. The Cybersecurity Law establishes for the first time in law a security assessment system for the outbound transfer of personal information and important data of critical information infrastructure operators and authorizes the national cybersecurity and informatization department to work with other regulatory authorities to formulate detailed security assessment implementation measures. In April 2017, the "Measures for Security Assessment of Personal Information and Important Data Transfer Abroad (Draft for Comments)" was released, establishing a basic framework for data transfer abroad. Subsequently, the "Data Transfer Security Assessment Guidelines (Draft)" and "Data Transfer Security Assessment Guidelines (Draft for Comments)" further specified the framework, clarified the concepts, and refined the security assessment process. In June 2019, the Cyberspace Administration of China released the "Measures for Security Assessment of Personal Information Transfer Abroad (Draft for Comments)", which details the assessment process for the transfer of personal information abroad to ensure the security of personal information in cross-border data flows.

It is worth noting that at this time, the regulations and national standards of various departments are mostly formulated based on the Cybersecurity Law, and some are formulated based on the National Security Law. However, at this time, China's data export regulations are largely departmental implemental rules detailing the above-mentioned laws. There is no high-level law to guide the system.

3. Attitude Changes and Three Pillars of Data Export Legislation

2021 is the year when data export legislation will be more perfect, and it will also be the year when the regulatory attitude in legislation shifts from data localization to a balance between data protection and utilization. In April, the State Council executive meeting passed the "Critical Information Infrastructure Security Protection Regulations" as administrative regulations. In June, the Data Security Law came into being. In this law, the legislative purpose is eye-catching. Among them, the legislative purpose of "promoting data development and utilization" appears for the first time, and precedes the statement of "protecting the legitimate rights and interests of individuals and organizations", implying that the country's regulatory attitude towards data export has begun to change, and it recognizes the importance of data development and utilization. Where necessity and value lie, the balance is quietly tilting from data localization to the orderly and free flow of data by the law. Article 3, Paragraph 3 of the "Data Security Law" directly states that it is necessary to ensure that data is in a state of effective protection and legal use, emphasizing the balance between protection and use. In August, the Standing Committee of the 13th National People's Congress passed the "Personal Information Protection Law", which provides a special chapter on cross-border rules for personal information and also mentions "promoting the reasonable use of personal information" in the legislative purpose. At the end of October, the Cyberspace Administration of China released the "Measures for Security Assessment of Data Transfer Abroad (Draft for Comments)", which further reflects the regulatory tendency of the free flow of data by the law. In November, the Cyberspace Administration of China issued the "Regulations on the Management of Network Data

Security (Draft for Comments)", which is a relatively high-level administrative regulation and is the same as the "Measures for the Security Assessment of Data Transfer Abroad (Draft for Comments)" which is a departmental regulation. , its nature is to refine and supplement the three-part superior method.

In July 2022, with the promulgation of the "Data Outbound Security Assessment Measures", the scope, conditions, and procedures of data outbound security assessment were specifically implemented, becoming a beacon in the sea of data outbound security assessment. Since then, mainland China's data export legislation has established clear-level and systematic data cross-border governance rules.

It is not difficult to find that the improvement of mainland China's data export legislation is reflected in the higher level of standards, wider coverage, stronger operability, and increased flow permissions, which is in sharp contrast to the early data export legislation.

C. Legal Principles and Reasons for Data Export Supervision

Data is the lifeblood of China's digital transformation and a strategic asset with very important strategic value. Regarding the regulatory legal basis for cross-border data flow, scholar Ding Xiaodong summarizes it into four categories: data without borders, data sovereignty, free trade of data, and cross-border human rights protection of data. This article will follow Ding Xiaodong's classification and further analyze this basis. The legal basis for cross-border travel from Mainland China to the Hong Kong Special Administrative Region.

1. Legal Basis for Cross-Border Data

a. Data Has No Borders and the Legal Principles of Data Sovereignty Do Not Apply

The two concepts of data without borders and data sovereignty do not apply to data export from mainland China to Hong Kong, China. As far as data without borders is concerned, because both mainland China and the Hong Kong Special Administrative Region belong to China, and are essentially data flows within the same country, the legal principle of data without borders does not apply to the situation discussed in this article. In my opinion, the legal theory of data sovereignty is very closely related to the practice of data localization. The concept of data sovereignty means that data should be subject to the laws and regulations of the nation-state where it is generated and processed, which is also a political effort to restrict data services across national borders.[1339] The basis of data sovereignty is the lack of an international legal framework for managing data. In this case, domestic policymakers in each country develop different systems of rules and processes to expand their jurisdictional control over the digital world domestically and internationally.[1340] However, in the context of "one country, two systems" and the Constitution as the fundamental law governing the

[1339] LIU L Z. The Rise of Data Politics: Digital China and the *World. Studies in Comparative International Development*, 2021, 56: 45–67.
[1340] Arner, Douglas W., Castellano, Giuliano G., & Selga, Eriks. The Transnational Data Governance Problem. *Berkeley Technology Law Journal*, 2021, 37(2): 623-699.

"Data Security Law", "Personal Information Protection Law", "Hong Kong Basic Law", etc., it is inappropriate for mainland China and the Hong Kong Special Administrative Region to establish data sovereignty separately. of. At the same time, China is also actively pursuing the "Beijing Effect" and seeking to expand its control and influence over data and data infrastructure globally.[1341] For the above reasons, China has no reason or need to establish two data sovereignty.

b. Data-free Trade and Data Cross-Border Human Rights Protection Cannot Be Transplanted

Free trade in data and cross-border human rights protection mainly reflect the attitudes of the United States and the European Union, and the methods adopted reflect specific cultural, political, economic, and legal characteristics.[1342] Historically, the United States has adopted a laissez-faire approach to data and technology. The complete transferability of data makes the property attribute of data more prominent and obvious. According to the blueprint provided by the "Washington Consensus", the development of the Internet tends to impose minimal regulation on data, creating a frictionless and pro-business environment for cross-border flows.[1343] It is under this model that the United States has given birth to the technology champions of Silicon Valley-Google, Apple, Facebook, Amazon, and Microsoft, becoming the country with the most big tech in the world. However, such a development model is believed by scholar Rogier Creemers to only exist in China before 2020. The emergence of a few large-scale dominant data companies is not what China currently wants to see, because the data outflow of these large-scale dominant data companies is likely to bring security threats in various aspects. The emergency suspension of Didi Chuxing's listing in the United States is very vivid. illustration.

The theory of human rights protection originated in the European Union and has influenced many other countries. Based on this theory, the legal theory of cross-border human rights protection of data has been derived. Underpinned by human rights, Data Governance also aims to embed a rights-based approach to data that reflects Europe's core cultural values and historical experience and to harmonize and extend consumer protection and data privacy across the 27 member states.[1344] However, the EU's approach is closely related to its actual situation. Internally, it established the General Data Protection Regulation and externally set data standards that can confront the United States and influence the world through the Brussels Effect, hoping to maintain its position, taking the line of coordinated market capitalism. The reality of the EU's coordination among its 27 member states is very different from the situation of coordination between mainland China and Hong Kong. China is a major player in digital technology and digital economy. Balancing rights protection and development interests is an urgent need for China. Hong Kong's economic position as a free trade center and constitutional position as a SAR in China give China an upper hand in

---

[1341] MATTHEW S E, THOMAS S. The Beijing Effect: China's 'Digital Silk Road' as Transnational Data Governance, *New York University Journal of International Law And Politics*, 2021, 54(1): 1-92.
[1342] FRANCESCA B, R DANIEL K. Kagan's Atlantic Crossing: Adversarial Legalism, Euro-legalism, and Cooperative Legalism. *GWU Law School Public Law Research Paper*, 2017, 66:1-27.
[1343] Arner, Douglas W., Castellano, Giuliano G., & Selga, Eriks. The Transnational Data Governance Problem. *Berkeley Technology Law Journal*, 2021, 37(2): 623-699.
[1344] ARMIN V B. The European Union as a Human Rights Organization? Human Rights and the Core of the European Union. *Common Market Law Review*, 2000, 37.

developing a more balanced data transfer regime.

c. Legal Basis for Data Governance in Mainland China

Mainland China's regulatory approach to digital competitiveness is characterized by "digital mercantilism" that focuses on ensuring economic stability.[1345] It is a style that revolves around property-based, rights-based, and state-centered data ownership and control. The Chinese data market is characterized by a combination of a property-based approach similar to that in the United States, in the context of private sector acquisition and control of data, and some form of restriction on the introduction of substitution from outside competition, with the government working closely with the non-state sector to reduce risks and achieve broader government objectives. China's regulatory approach is a "unique combination of data protection and government control of data flows", embodying the state-centered approach to ensuring data security.

2. Reasons for Data Export Supervision

Data does not generate or appear alone, and the individuals, companies, and countries behind the data are all separated by national boundaries. The protection of personal rights and interests, property attributes, and national security attached to data need to be regulated. If left unregulated, large multinational corporations will directly dominate the cross-border transmission of data and have power beyond sovereignty. The volume of data exported from mainland companies listed in Hong Kong should not be underestimated. Even if Mainland China and Hong Kong are under the same Chinese sovereignty, data flow from the Mainland to Hong Kong still needs to be regulated.

The first reason is data security considerations. Article 3 of Mainland China's "Data Security Law" stipulates that data security refers to taking necessary measures to ensure that data is in a state of effective protection and legal utilization, as well as the ability to ensure continued security. Article 76, Paragraph 2, of China's Cybersecurity Law, defines "data security" as "the ability to ensure the integrity, confidentiality, and availability of network data." "Confidentiality" here means that the data cannot be obtained by others who should not have access; "integrity" means that the data is not tampered with without authorization or can be quickly discovered after tampering; "availability" means that the data meets the requirements of consistency, accuracy, Timeliness requirements. Even if the data flows from the mainland to Hong Kong rather than abroad, issues of confidentiality and integrity still exist.

The second is out of consideration for protecting personal information. The data export risk management system originated from the protection of personal rights and interests under the cross-border flow of information. Personal information is the most common type of data subject to localized storage requirements, and it is also a category for which mainland China has clarified localized storage of data in its early legislation. Data export may include the transfer of sensitive data such as medical care, health, bank account passwords, genetic information, etc. If not supervised and controlled, personal and property safety information will be leaked, leading to the risk of infringement.

[1345] CYRUS C, PO-CHING L. E-Commerce Mercantilism-Practices and Causes. *Journal Of International Trade Law And Policy*, 2020.

The third is due to national security considerations. In addition to the transfer of sensitive personal data, data export will also include some data containing the economic performance and trends of enterprises and even countries, such as government procurement data, important economic data, involving specific aspects of national politics, society, and economy, and even May contain military data, etc. The importance of these data is self-evident. Due to national security considerations, the aforementioned data should be restricted from exporting abroad, so there is an even greater need to supervise exported data.

The fourth is due to considerations of other legitimate public policy objectives. Since data has property attributes, the data itself can bring economic benefits to the enterprise. If companies are worried about the creative destruction caused by data sharing, companies can hoard data for themselves, but this will limit the positive externalities and welfare benefits that data can generate; Insufficient investment in privacy protection will amplify the negative externalities of data. As noted economist Daron Acemoglu has argued, Big Tech's most pernicious impact stems from their ability to direct technological change, since these companies only have incentives to fund projects that are compatible with their own interests and business models. Research. Therefore, data must not only flow but also circulate on a safe and orderly basis. Risk control of outbound flows is inseparable from outbound supervision.

In summary, it is still necessary and important to supervise the export of data from the mainland to Hong Kong.

D. Data Export: Standards for Data Flow from Mainland China to Hong Kong

The most typical situation where data from mainland China flows to the Hong Kong Special Administrative Region is when companies go public in Hong Kong. As mentioned in the previous introduction, on the one hand, due to the strengthening of cybersecurity supervision of overseas listed companies and the intensification of regulatory friction between China and the United States, there are many obstacles for Chinese companies to list in the United States. The cybersecurity reviews regulatory requirements introduced by mainland China are for those seeking to enter the United States. Chinese companies in the capital market have increased additional transaction costs; on the other hand, the supportive attitude of mainland regulatory agencies towards listing in Hong Kong and Hong Kong's status as a financial center have added confidence and protection to companies listing in Hong Kong. Hong Kong has become China's first choice for corporate IPOs. At present, mainland China has a basic regulatory framework for the rules on the export of personal information, and several departmental laws provide systematic provisions. However, there are still some problems with the lack of specific details and ambiguity of some rules. This part will be elaborated by citing the legal text.

1. Personal Information Export under the "Personal Information Protection Law"

The "Personal Information Protection Law" stipulates five requirements for the transfer of personal information abroad, and specifically formulates rules for the cross-

border provision of personal information in Chapter 3. It is worth pointing out that through the provisions of Article 40 of the "Personal Information Protection Law" plus the exceptions, it can be seen that after reaching a certain level of information importance and information quantity, data export is an exception, that is, in principle, It should be stored within the country and only provided abroad when it is necessary. Only then can information be exported, and many requirements must be met. Article 38 sets out the prerequisites for the export of personal information data abroad, which is the first requirement. This requirement seems to be elaborated in the article by enumerating and adding redundant clauses, which seems to be a relatively comprehensive expression. However, Article 38 (1) mentions "by the provisions of Article 40" The security assessment carried out remains very vague in Article 40. Article 40 of the "Personal Information Protection Law" that came into effect on November 1, 2021, simply outlines the objective conditions of the security assessment. What is the content of the security assessment and what are the criteria for assessment are not specified? However, three days before the "Personal Information Protection Law" came into effect, the Cyberspace Administration of China released the "Data Transfer Security Assessment Measures (Draft for Comments)" to solicit public opinion. In less than a year, the Cyberspace Administration of China issued the "Measures for Data Exit Security Assessment" as departmental regulations. This measure also regulates the subject of data processors as well as the departments, procedures, assessment matters, materials, etc. that apply to data export security assessment. Detailed regulations have been made, stipulating that data processors that process the personal information of more than 1 million people and data processors that have provided personal information of 100,000 people or sensitive personal information of 10,000 people overseas since January 1 last year need to declare. For personal information processors who enter into contracts with overseas recipients by Article 38, paragraph 1, item (3) of the Personal Information Protection Law, there are also "Standard Contract Regulations for the Transfer of Personal Information Abroad" issued in June 2022 (Draft for comments)" can be referred to.

The second and third requirements for the export of personal information are notification and individual consent, which are stipulated in Article 39 of the Personal Information Protection Law. Notification is the special notification obligation of personal information processors in addition to the constraints stipulated in Article 17 of the "Personal Information Protection Law". It is a manifestation of protecting the information subject's right to know, including but not limited to name, contact information, processing purpose, and processing method. etc.; individual consent is "individual consent" in special circumstances, and is a manifestation of the individual's ability to exercise decision-making power.

The fourth and fifth requirements for the export of personal information are personal information protection impact assessment and safeguarding recipient standards. The fourth element is expressly stipulated in Articles 55 and 56 of the Personal Information Protection Law, which adopts an evaluation method similar to the principle of proportionality. The requirement of ensuring recipient standards is stipulated in Article 38, Paragraph 2 of the Personal Information Protection Law, which is also a requirement that this article considers to be very important. When citizens' data flows to jurisdictions that do not provide them with a comparable level of privacy protection, such transfers may undermine privacy objectives, and this concern may further motivate outbound regulators to restrict the free flow of data across borders. The

receiving party's information protection standards largely determine the data exporting party's review attitude toward whether the data can be exported. It is also a solid guarantee for the personal rights and interests behind the data. After all, the first legislative purpose of the "Personal Information Protection Law" The first is "protecting personal information rights and interests", followed by "regulating personal information processing activities" and "promoting the reasonable use of personal information."

To sum up, to protect personal information stored in the Mainland, the state's supervision of personal information data is no different whether it flows from the Mainland to Hong Kong or from the Mainland to abroad. When the amount of information collected and generated by personal information processors in the mainland reaches a certain amount, the regulatory red line of "in principle, the information needs to be stored within the territory" is triggered. If it is really necessary to leave the country, the above five requirements must be met. On the contrary, if it does not reach a certain level, personal information can be provided overseas if any one of the conditions stipulated in Article 38 is met, and the data can be exported abroad.

2. Transfer of Important Information Abroad under the "Cybersecurity Law" and "Data Security Law"

In the "Data Transfer Security Assessment Guide (Draft)", the determination of whether data is important is based on the combination of national security, economic development, and social and public interests, and twenty-eight categories are listed in the appendix. The "Measures for Security Assessment of Data Transfer Abroad" also combines the above three aspects to define important data, but refines social public interests into social stability, public health, and safety. It can be seen that restrictions on the export of important data are largely based on considerations of national security rather than individual rights, and are completely different from the "protection of personal information rights and interests" that is most important in the regulation of the export of personal information.

Article 31 of the "Data Security Law" and Article 37 of the "Cybersecurity Law" both provide for the export of important information abroad. Article 31 of the "Data Security Law" classifies important data according to the holding entities, and stipulates that if important data needs to be provided overseas, a security assessment must be conducted by the methods formulated by the national cybersecurity and informatization department in conjunction with relevant departments of the State Council. It can be seen that even if important information and data flow from the mainland to Hong Kong, data outbound security assessment is an essential link. This article believes that adopting such an attitude is also related to the lack of a framework for data export supervision in the Hong Kong Special Administrative Region. This part will be discussed in Chapter 2 of this article.

It is worth pointing out that Article 13 of the "Network Data Security Management Regulations (Draft for Comments)" considers the data export situation involved in mainland companies listing in Hong Kong separately, and does not review all companies listed in Hong Kong across the board. From the perspective of contextual interpretation, companies listed in Hong Kong that "may affect national security" are

subject to the same cybersecurity review requirements as data processors listed abroad that handle the personal information of more than one million people. Combined with the fourth safety clause, it is not difficult to find that data processors listed in Hong Kong only use "may affect national security" as the premise for declaration because "may affect national security" is the real reason and starting point for supervision. For data processors that do not involve the country, Safe data processors go from the mainland to Hong Kong, which is the so-called overseas listing. The mainland government does not have a strong will to restrict or supervise.

To sum up, when it comes to exporting important data abroad, legislators' main concern is national security. If the data flow from the mainland to Hong Kong does not involve national security and is just a commercial activity, legislators believe that there is no need to interfere too much. However, if the data flow from the mainland to Hong Kong may affect national security, it needs to be subject to the same strict review and supervision as the data flow abroad.

It is worth noting that network security review and data export security assessment may overlap during the operation of the system. For example, if a data-based enterprise has a large amount of personal information in its operations and involves cross-border transmission of data, cross-application will occur. In the absence of a clear applicable relationship between the two in existing regulations, the coordination and connection between the network security review system and the data export security assessment system still need to be further clarified.

II. REGULATION OF MAINLAND DATA RETRIEVAL IN HONG KONG, CHINA

A. Hong Kong’s Regulatory Orientation and Development Path for Data Protection

The Hong Kong Special Administrative Region is both an international financial center and the center of massive data inflows and outflows. At the same time, because Hong Kong has a highly transparent regulatory system and a laissez-faire business model, as a major international business center, there are more than 4,000 regional headquarters and offices of leading multinational companies in Hong Kong, attracting a large international data flow.[1346]

Hong Kong's attitude towards data flow has also attracted more cross-border data inflows, forming a positive cycle of "data inflow – free supervision – data inflow – free supervision". This cycle can be specifically reflected in the following: Hong Kong, based on its status as a commercial and financial center, will generate a large amount of inbound and outbound data flows. These cross-border data flows can reduce information asymmetry and improve market access opportunities. They can not only be used for business transactions, supply chain forecasting, market access, and customs processing but can also support cross-border operations necessary for cross-border agreements, ownership of key logistics facilities, and physical and digital delivery of

[1346] Hong Kong Government. LCQ7: Foreign Companies’ Regional Headquarters and Offices in Hong Kong. 2021-02-24.https://www.info.gov.hk/gia/general/202102/24/P2021022400302p.htm.

goods and services, bringing benefits and benefits to Hong Kong, and to some extent enhance the city's competitiveness.[1347] Hong Kong, which has enjoyed the dividends brought by data flows, has also been acquiescing to the fact that the law on data export, namely Article 33 of the Personal Data (Privacy) Ordinance (PDPO), has not come into effect. It acquiesces to the prohibition of data localization requirements and is relatively free in supervising data exports. Therefore, Hong Kong's tolerance further attracts the inflow of overseas data.

It can be said that a large amount of data flow is not only the inevitable result of Hong Kong itself as a financial center, but also a means for Hong Kong to further strengthen its position as an international business city and international technology city. As Arner said, the characteristics of each jurisdiction are based on its attitude towards the market and governance, the normative principles that support the exercise of control over data, and the mode of regulating data, and the constantly evolving and unique data governance style.[1348] Hong Kong SAR's current attitude towards cross-border flows is more similar to that of the United States. Compared with the protection of rights and interests, the Hong Kong SAR pays more attention to the commercial benefits brought by the free flow of data.

B. Hong Kong's General Data Protection Model as Reflected in the Personal Data (Privacy) Ordinance

China's Hong Kong Special Administrative Region is the first jurisdiction in Asia to enact comprehensive protection of personal data privacy, and the protection of personal information has always been among the best. As early as 1995, the Hong Kong Special Administrative Region enacted the Personal Data (Privacy) Ordinance. The Office of the Privacy Commissioner has also formulated and issued a series of "New Guidelines on Direct Marketing", "Guidelines on Cross-border Data Transfers", "Guidelines on the Collection and Use of Biometric Data", "Guidelines for Employers and Human Resources Managers" and "Best Practice Guidelines for Mobile Application Development". Guidelines such as "Instructions for Employers to Supervise Employee Work Activities" and "Code of Practice on Identity Card Numbers and Other Identity Codes" help data users understand the relevant provisions of the "Privacy Ordinance" more clearly. In addition, the "Personal Data (Privacy) Ordinance" has been passed many times After comprehensive review and consultation revision, it can be said that the Hong Kong SAR has a relatively complete personal data protection system.

In terms of regulatory entities, the Hong Kong SAR has established the Hong Kong Office of the Privacy Commissioner for Personal Data (PCPD), which monitors and supervises compliance and implementation by issuing guidelines and other measures. The values pursued by the PCPD are respected (respecting the personal data privacy of others), integrity (acting fairly and professionally), innovation (keeping up with technological, social, and economic developments), independence (independence

---

[1347] Hong Kong International Airport.Hong Kong and Shenzhen Airports Sign Cooperation Agreement Join Hands to Promote Airspace Resources Optimisation. 2021-01-04. https://www.hongkongairport.com/en/media-centre/press-release/2016/pr_1200.

[1348] Arner, Douglas W., Castellano, Giuliano G., & Selga, Eriks. The Transnational Data Governance Problem. *Berkeley Technology Law Journal*, 2021, 37(2): 623-699.

from the government and other institutions), and excellence (committed to pursuing Best results and highest standards). This is also mutually confirmed with the legislative purpose of Hong Kong's Personal Data (Privacy) Ordinance, which is "the purpose of the Ordinance is to protect the right to privacy about personal data…". According to the Ordinance, individuals are granted 10 main rights: the right to provide only the required information; the right to collect the information fairly and for lawful purposes; the right to be informed of the purpose of the data; the right to require that the data be accurate; and the right to require that the data not be excessively retained. Rights; the right to refuse consent to change of data use; the right to require data security measures; the right to be informed of data policies and measures; the right to access data; the right to correct data. These ten rights are basic rights related to the protection of personal information, and most of them provide for minimum provision, anytime access and modification, maximum access to purposes, and high convenience for removal.

Combining Hong Kong's Personal Data (Privacy) Ordinance and the value pursuit of the Office of the Privacy Commissioner for Personal Data (PCPD), Hong Kong pays special attention to the protection of privacy when it comes to personal data. The establishment of the Ordinance and the PCPD will protect and respect Personal privacy as the primary pursuit.

C. Hong Kong's Data Re-export Framework

No matter which country or region data is exported to, there is a possibility of secondary export in that country or region, and data exported from mainland China to the Hong Kong SAR is no exception. Therefore, it is necessary to explore Hong Kong's data re-export framework to anticipate possible risks after data is exported abroad.

Hong Kong's "Personal Data (Privacy) Ordinance" takes the protection of personal information and personal privacy as its main legislative tendency and is a conservative protection of personal privacy. According to the provisions of Article 33, paragraph 1, of the Ordinance, the export of data from mainland China After entering the Hong Kong SAR, this batch of data meets the subject requirements of Article 33, paragraph 1, and should also be subject to restrictions on re-exit. Except for the circumstances listed in paragraph 2, it is generally not allowed to re-exit the country. However, the provisions of Article 33 of the Personal Data (Privacy) Ordinance regarding data export abroad have never been implemented, resulting in the Ordinance appearing to be "very generous" in terms of data export.

The debate on whether Article 33 should be implemented in the Hong Kong Special Administrative Region has never ceased, and the minds of legislators are constantly swinging. Hong Kong hopes to participate in the development of global data policies and further promote itself as a sound and stable business place.[1349] In January 2020, the Legislative Council Committee on Constitutional Affairs debated proposed changes to the Personal Data (Privacy) Ordinance. Including mandatory data breach notification mechanisms and data retention periods, and stated that templates and best practice guidelines related to cross-border transfers between institutions and cross-border transfers between cloud processors should be issued; but at the same time, Hong

[1349] Legislative Council Hong Kong.Review of the Personal Data (Privacy) Ordinance. 2020-01-20. https://www. legco.gov.hk/yr19-20/english/panels/ca/papers/ca20200120cb2-512-4-e.pdf.

Kong cannot resist prohibiting data localization on Benefits of own city development competitiveness, prohibiting data localization means allowing data outsourcing, which can reduce the cost of international business and promote openness for multinational companies headquartered in Hong Kong and data-driven enterprises engaged in finance, logistics and innovation and a flexible global technology infrastructure are critical.[1350] Finally, in the amendments to the Personal Data (Privacy) Ordinance published in July 2020, the focus is on combating leaks rather than cross-border data flows. Article 33 on data localization is still pending.

The fact that Article 33 has come into effect does not mean that there are no restrictions on the export of data after it has been transferred to the Hong Kong SAR. The absence of cross-border data restrictions does not mean that users are free to transfer data outside the jurisdiction, as users remain ultimately responsible for their data and are subject to the data protection principles of the Personal Data (Privacy) Ordinance. At the same time, the Office of the Privacy Commissioner for Personal Data in Hong Kong has also issued cross-border data transfer guidelines to help data users understand the requirements after Article 33 comes into effect and provide practical guidance. The PCPD also seeks to encourage data users to adopt the practices recommended in the Guidelines as part of their corporate governance responsibilities.[1351]

It is worth mentioning that Hong Kong's current approach to managing cross-border data flows is consistent with its business approach – it is built on a patchwork of legislation and "default" policies promised by free trade agreements.[1352] Although Hong Kong does not have a clear strategy, policy, or rules to manage data entering and leaving Hong Kong, the Hong Kong government adopts a tacit policy, and individual issues can find guidance in laws and regulations such as the Personal Data (Privacy) Ordinance, and free trade agreements. The commitments made by Hong Kong are also a considerable complement to Hong Kong's efforts in managing cross-border data flows. Hong Kong's commitments in the ASEAN-Hong Kong Free Trade Agreement (AHKFTA Agreement) the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP) and the United States-Mexico-Canada Agreement (USMCA) regarding the free flow of data and prohibition of data localization The promises are similar. By making these commitments, Hong Kong strengthens its ambition to become a global center for receiving, storing, and sharing finance and data, and more broadly demonstrates its goal of maintaining free cross-border data flows.

III. PRACTICAL PROBLEMS IN THE FLOW OF DATA FROM MAINLAND CHINA TO THE HONG KONG SPECIAL ADMINISTRATIVE REGION

[1350] CORY, NIGEL. Cross-Border Data Flows: Where are the Barriers, and What do They Cost?. 2017-05-01. https://itif.org/publications/2017/05/01/cross-border-data-flows-where-are-barriers-and-what-do-they-cost.

[1351] Hong Kong Privacy Commissioner for Personal Data. Guidance on Personal Data Protection in Cross-border Data Transfer. 2017-11-10). https://www.pcpd.org.hk/english/resources_centre/publications/files/GN_ crossborder_e.pdf.

[1352] Mercurio, Bryan. On the Importance of Developing a Coherent Policy Facilitating and Regulating Cross-Border Data Flows. *International Trade Law and Regulation*, 2021(1): 97–104.

A. The Pull between the Development Orientations of the Two Places: Rights or Development?

As mentioned above, due to the needs of data development or the different stages of data development, the legislation and policy trends on data protection and export in mainland China and Hong Kong SAR are not completely consistent, resulting in different data governance styles. Relatively speaking, the Hong Kong Special Administrative Region is more open and mobile and pays more attention to regional development in rights protection and regional development, while Mainland China attaches more importance to localized storage and strives to find stable development and stable protection in rights protection and regional development. balance.

Before 2009, the Chinese government largely followed U.S. practices regarding private data within China, but over the past decade, the Chinese government has tightened controls on the flow of data in and out of China, controlling access to data at home and abroad, Monitoring, regulating and controlling the monopoly power created by the concentration of data sent or collected by large technology companies, eventually almost eliminating large IT companies such as Google, Apple, Meta, Amazon, Microsoft (GAFAM) from the domestic market, hoping to cultivate A competitive local leading enterprise. In the process of controlling the expansion of large IT companies, the Chinese government will inevitably have some impact on the Hong Kong Special Administrative Region, which aspires to economic development. However, data shows that mainland China's target of large technology companies has temporarily affected the stock prices of these companies on local exchanges. It only affected Hong Kong indirectly, but many experts are still worried that the impact may be more direct in the future.

Although mainland China's legislative provisions only apply to all companies operating in China, laws and regulations such as the "Personal Information Protection Law" and "Data Export Security Assessment Measures" cannot be applied to the Hong Kong SAR, these provisions may still have an impact on Hong Kong. The special zone has a significant impact. Among the new rules introduced by Beijing in the past five years, companies are required to obtain government approval before transferring certain types of data outside China, as part of a broader government effort to tighten controls over data and protect national security. Hong Kong has long been a hub for international companies to enter the mainland market. Many multinational companies have established regional headquarters in Hong Kong, hoping to ride on Hong Kong's proximity to mainland China and its status as a global financial center. The promulgation of these new regulations may prevent these multinational companies from storing sensitive data in Hong Kong, making it more difficult for these multinational companies to operate in Hong Kong, China.

This is evident from the pull between the development orientations of the two places. Hong Kong's economy is highly dependent on its role as the world's gateway to mainland China. According to a report by the Hong Kong Trade Development Council, about 60% of Hong Kong's exports are re-exports - goods imported into Hong Kong from countries around the world, which are then re-exported to other markets, including goods entering mainland China. If foreign companies are forced to store data on the mainland, this could undermine Hong Kong's status as a center for import and export

transactions, as it will not be able to offer the same level of data privacy and security as other global financial centers. Hong Kong's status as a global financial center may also be affected. If multinational banks and financial institutions that set up operations in Hong Kong realize that Hong Kong can no longer provide them with a good business environment and the costs of operations and compliance have greatly increased, they are likely to Move to other financial centers in Asia, such as Singapore.

In general, the promulgation of new regulations on the mainland in the past five years has posed new challenges to Hong Kong's development. However, this article believes that new challenges do not mean that they will bring new disadvantages to the Hong Kong region, or hinder the development of the Hong Kong region. The mainland and Hong Kong are each part of China. Although the systems are different, they should both make efforts for data control and maintain National Security. However, how to achieve common development still requires the mainland and Hong Kong to think and work together.

B. Tightening of Legal Regulations Between the Two Places: Hong Kong Has Become a Shortcoming in Data Export

Since the laws and regulations of mainland China and the laws and regulations of Hong Kong are at different times, have different levels of development, and have different tightness of regulations, whether the two places are fully aligned in terms of data export and whether there is an overlap or vacuum in supervision is something that this article believes needs to be discussed. The author compared Hong Kong's Personal Data (Privacy) Ordinance and Mainland China's Personal Information Protection Law in terms of legislative purposes, the definition of "personal data", consent models, fines, corporate compliance qualifications, and codes of conduct. In terms of legislative purposes, it is not difficult to find that Hong Kong's regulations do not clearly state "promoting the reasonable use of personal information", but it is mentioned in the Mainland's "Personal Information Protection Law". This does not mean that personal information in Hong Kong currently adopts a conservative approach to rights protection, but because Hong Kong's Personal Data (Privacy) Ordinance has been enacted for a long time and has not been revised and updated promptly for legislative purposes. This also reflects that because Hong Kong enacted regulations earlier, with the development of the times and technology, there is some lag, which may create loopholes in the protection of personal data.

In terms of the definition of personal data, Hong Kong stipulates that the subject of personal data must be a living individual, while the Mainland's "Personal Information Protection Law" stipulates that the subject of personal information only requires natural persons, and does not require the natural person to be alive. The first problem that may arise from this is that during the process of data export, if the personal information of a deceased person is transmitted, the protection that can be obtained in the mainland is not available in the Hong Kong SAR, resulting in data protection. on the fault. In addition, Hong Kong does not specifically distinguish between personal information and personal sensitive information, while the mainland classifies personal data in regulations such as the "Guidelines for Security Assessment of Information Security Technology Data Transfer (Draft)" and stipulates personal information and personal sensitive information. Different types of information are given different levels of protection, which will lead to the problem that data from mainland China can no longer

receive the same level of protection after being exported abroad.

In terms of penalties, according to Hong Kong laws and regulations, violations of the regulations can be punished with a maximum penalty of HK$1 million (approximately 880,000 yuan) and five years of imprisonment. According to mainland regulations, a maximum penalty of not more than 50 million yuan or a turnover of 100 million yuan in the previous year can also be imposed. A fine of not more than five-fifths of the amount shall be imposed, and a fine of not less than RMB 100,000 but not more than RMB 1,000,000 shall be imposed on the directly responsible person in charge and other directly responsible personnel. Judging from the intensity of fines, the penalties in the Mainland are even higher, especially for enterprises, especially large enterprises. The turnover of less than 5% of the previous year is likely to be much higher than the one million Hong Kong dollar cap stipulated in Hong Kong. In terms of the objects of punishment, Hong Kong's current laws only stipulate personal liability, while mainland laws distinguish between individual liability and corporate liability. Enterprises not only face high financial penalties, but also may be ordered to suspend relevant business or suspend business for rectification or notify Relevant competent authorities will revoke relevant business licenses or revoke business licenses, and be subject to administrative penalties. Today's enterprises are increasingly adopting data-driven business models and strategies to gain and sustain a competitive "data advantage" over their opponents. Generally speaking, a reduction in the cost of violating the law is likely to lead to an increase in violations. Under the economic thinking model of cost-benefit analysis, when the cost of violating the law is too low and is lower than the benefits obtained from data flow, companies as data controllers are likely to take risks and would rather pay fines than use Hong Kong to complete data entry and exit.

Generally speaking, the degree of tightness of legal regulations in the two places is different. Hong Kong's relatively free and relaxed regulations make Hong Kong likely to become a shortcoming in data export, overriding the protection of data under mainland China's laws and regulations, and causing many practical problems.

IV. POSSIBLE LEGAL SOLUTIONS FOR DATA FLOW FROM MAINLAND CHINA TO THE HONG KONG SAR

Huang Ning and Li Yang pointed out that there are three difficulties in the regulation of cross-border data flows, that is, "good data protection", "free flow of cross-border data" and "data protection autonomy" of various governments cannot be achieved at the same time[1353]; D. W. Arner pointed out that the domestic governance styles of each country are consolidated into competing and conflicting data governance systems, the transnational output and influence of various countries are destroying the existing transnational data governance paradigm based on the free flow of data and hindering international coordination in the global data economy. [1354] This is transnational data. the wicked problem of transnational data governance because there

[1353] 黄宁,李杨."三难选择"下跨境数据流动规制的演进与成因.清华大学学报(哲学社会科学版),2017,32(05):172-182+199.

[1354] Arner, Douglas W., Castellano, Giuliano G., & Selga, Eriks. The Transnational Data Governance Problem. *Berkeley Technology Law Journal*, 2021, 37(2): 623-699.

is no single solution to it.[1355] It can be said that data export cannot take into account both the protection and flow needs of data at the transnational level. The balance will inevitably be more or less tilted towards a certain value. However, this article believes that it contains the possible laws of data flow from mainland China to the Hong Kong Special Administrative Region. solution. The reason why data governance mechanisms are different is due to conflicts of national interests. However, it can be seen that there is no conflict of national interests between mainland China and Hong Kong, and it is entirely possible to achieve win-win results.

A. Jointly Negotiate to Establish Data Circulation and Transaction Standards Suitable for the Characteristics of Greater China

Various laws and regulations promulgated by mainland China in the past five years have indirectly affected Hong Kong, and at the same time, they have also brought new opportunities and challenges to Hong Kong. Under the basic framework of one country, and two systems, the mainland and Hong Kong can jointly negotiate and establish data circulation and transaction standards that suit the characteristics of Greater China.

Allowing some freedom for cross-border data flows would benefit both Hong Kong and the mainland. With more than 700 million internet users across China, leading technology manufacturing companies such as Huawei and Lenovo, and growing technology giants such as Alibaba, Baidu, and Tencent, allowing partial freedom in the flow of data across borders would not only allow Hong Kong to maintain its position among multinational companies The competitiveness in mind will also bring obvious economic benefits to the mainland due to the cross-border flow of data. In addition, China's local technology companies also have plans and ambitions to enter the global market. If too many restrictions are imposed on cross-border data flows, it may hinder the development of these companies' global operations and reduce their competitiveness. At the same time, controlled freedom in cross-border data flows can also prompt Hong Kong to strengthen its legal framework for data protection and privacy. This will not only attract companies and maintain its status as a global financial center but also help Hong Kong in building "Asia's largest financial center". One of the "Secure Data Center Cities".

Hong Kong and the Mainland should jointly negotiate to establish data circulation and transaction standards suitable for the characteristics of Greater China, and discuss how to strike an appropriate balance in supervision while providing appropriate deterrence without compromising the interests of innovation, collaboration, and improving business efficiency. In the "14th Five-Year Plan" announced in March 2021, the mainland has established the goal of developing an innovative country and a technological power, and proposed a development pattern of "domestic large cycle and domestic and international dual cycle" to develop the Guangdong-Hong Kong-Macao Greater Bay Area. Become an international science and technology innovation center. Hong Kong also stated in the "Hong Kong Innovation and Technology Development Blueprint" that Goal 01 is to promote the effective flow of innovation elements across borders, strengthen the competitiveness of Hong Kong's innovation and technology,

---

[1355] Pedch U, E., Vermass P. The Wickedness of Rittel and Webber's Dilemmas. *Administration and Society*, 2020,52:960.

and better serve the needs of the country. Measures to achieve the goal include exploring with mainland ministries and commissions Implement more measures to promote the convenient cross-border flow of innovative elements. In terms of data, we will actively study with the mainland on specific facilitation arrangements to promote the flow of data from the mainland to Hong Kong and launch a pilot plan for cross-border data flow in the Greater Bay Area in 2023 to test Technical standards, measures and data governance mechanisms for widespread implementation in the future. Both sides have shown an attitude of active consultation, cooperation, and putting national interests first. If a data exchange agreement mechanism between mainland China and Hong Kong, or even Hong Kong, Macao, and Taiwan can be established shortly, it will be of great benefit to the development of both places.

B. Hong Kong Promotes Article 33 of the Ordinance to Take Effect as Soon as Possible to Fill the Shortcomings of Data Export Regime

Hong Kong hopes to consolidate its advantages as a data center in the Asia-Pacific region, not by becoming an "outbound paradise", but by promoting Article 33 of the Ordinance to take effect as soon as possible. Mainland China's laws have tended to regulate data exported to the Hong Kong Special Administrative Region and exported to foreign countries separately, because China is well aware that Hong Kong is still a part of China, and there is no national-level political and economic conflict between the mainland and Hong Kong. It can be said that All data is transferred between people. At present, the "Regulations (Draft for Comments)" have set different application conditions for data processors listed abroad and listed in Hong Kong, China. It is foreseeable that after the mainland laws and regulations are improved, the export of data to Hong Kong should be more relaxed than the export of data. To design abroad. To comprehensively regulate and manage privacy, Hong Kong should implement or amend Article 33 of the Ordinance to fill the shortcomings of data export so that Article 33 will no longer become a "backdoor shortcut" for multinational companies to enter the Chinese market.

C. Legal and Administrative Integration Between Hong Kong and the Mainland

It is recommended that Hong Kong and the Mainland align themselves on the administrative management of data export abroad. The author believes that although the EU is an attempt to establish a single market between countries and is somewhat different from China's national conditions of one country, two systems, some of the EU's practices are worth learning from. To manage data in a single unit, the EU launched the European Cloud initiative to simplify data access by seamlessly moving, sharing, and reusing data across European markets and borders. The EU is also creating its own data walled garden, designed to connect cloud providers across Europe, harmonize technology standards, and ensure data privacy and security walls. The author believes that we might as well set up a supervisory and management agency for data transmission from mainland China to the Hong Kong SAR. Currently, the cross-border data flow between the Mainland and the Hong Kong SAR is managed separately by different agencies on both sides. This will, to some extent, lead to the lack of an effective unified management mechanism and hinder cross-border data flow.

It is recommended that Hong Kong and the Mainland align with the legal

regulations on data export abroad. The Organization for Economic Co-operation and Development (OECD) advocates a more detailed differentiation of personal data based on international standards and divides technologies into five categories of data identification: (1) identifying data, (2) pseudonymous data, (3) unlinked pseudonyms Data, (4) Anonymous Data and (5) Aggregated Data. At present, mainland laws distinguish personal data, but it is not as detailed as advocated by the Organization for Economic Cooperation and Development. It may be further refined in the future. However, Hong Kong law has not yet distinguished personal data, and it needs to be revised and improved. The current development of data analytics and artificial intelligence has made it easier to link seemingly non-personal data with identified or identifiable individuals. The concept of "personal data" is constantly changing and is no longer as stable as it was when the Ordinance was originally formulated. definition, Hong Kong should take advantage of the opportunity to update the existing structure. Because a higher level of protection of personal data rights will also enhance consumer confidence in digital trade and stimulate economic development, the mainland and Hong Kong may wish to align on the legal regulations on data export, learn from each other's strengths, and eliminate the shortcomings of national security loopholes in data. plate. Before the legal integration, if a certain number of companies can achieve a high level of privacy protection, they can also learn from the "safe harbor" model and have the governments of the two places negotiate and sign bilateral data flow agreements.

D. Special Legal Arrangement for Data Transfer within the Guangdong-Hong Kong-Macao Greater Bay Area (GBA)

In December 2023, the Cybersecurity Administration of China (CAC), China's top cybersecurity authority, released a new set of guidelines for companies in the Guangdong-Hong Kong-Macao Greater Bay Area (GBA) to sign a standard contract to engage in cross-border personal information (PI) transfer between the mainland portion of the GBA and Hong Kong. The GBA (Mainland, Hong Kong) Implementation Guidelines for the Standard Contract for Cross-border Flow of Personal Information (the "GBA guidelines") are the result of an agreement between the CAC and the Innovation, Technology, and Industry Bureau (ITIB) of Hong Kong to facilitate cross-border data flows and establish security rules for PI transfer within the GBA. The GBA guidelines, which took effect on December 13, 2023, make it significantly easier for companies located in one of the nine mainland cities of the GBA to transfer personal information to Hong Kong by expanding the scope of companies permitted to use the standard contract procedure, as well as simplifying filing procedures. The efforts to streamline cross-border PI transfer align with the central goal of deepening integration between the mainland and offshore areas of the GBA and fostering a more business-friendly environment in the region. This is a positive development towards a sound legal framework for data transfer between the Mainland and Hong Kong.

CONCLUSION

The constitutional basis for the establishment of the Hong Kong Special Administrative Region is Article 31 of the Constitution. As the fundamental law, Article 31 of the Constitution constitutes the legal basis of "one country". Since Hong Kong returned to the motherland in 1997, it has enjoyed a "high degree of autonomy" under the authorization of the central government. The opportunities brought by "two systems" have made Hong Kong an important link between the mainland market and

the international market. Under the Constitution, the relationship between the mainland and Hong Kong follows the principle of "one country, two systems". Within the framework of one China, the mainland and Hong Kong maintain their respective systems and development models. Therefore, although data flows from the mainland to Hong Kong, China, it is data outbound, but the data has never left the country. The transmission of data from the mainland to Hong Kong essentially still belongs to the free flow of data within a country. Problems that may arise during the flow of data can be solved through the improvement of internal laws, the connection of internal administration, and the standardization of internal systems.

In an era when the value-generating function of data is growing, we should make good use of the convenience and advantages of "one country" and "two systems", seize opportunities, use the value of data to achieve national development and protect the rights of the people while developing. There is reason to believe that with the continuous improvement and integration of laws, the continued in-depth cooperation between the two places, and the continuous advancement of technology, the cross-border data flow from mainland China to the Hong Kong SAR will be safer, more reliable and more convenient in the future. China The mainland and Hong Kong as a whole will stand on the international stage and make greater contributions to the development of the global digitalization process.

This page intentionally left blank.

www.ijlet.org

La Nouvelle Jeunesse

ISSN 2769-7142

9 772769 714009