VOL 2024 NO 3 | SUMMER 2024

International Journal of Law, Ethics, and Technology



La Nouvelle Jeunesse

SUMMER 2024 The International Journal of Law, Ethics, and Technology Staff

Editor-in-Chief
George G. Zheng Shanghai Jiao Tong University, China

Associate Editors
Yan Pan Shanghai Jiao Tong University, China
You Zhang Huazhong University of Science and Technology, China
L. Ben Shanghai Jiao Tong University, China

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: August 30, 2024

Copyright © 2024 by La Nouvelle Jeunesse, except where otherwise indicated.

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 per the approval by the editors. To request a place on the list, please email us at info@ijlet.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

Can AI Make a Case? AI vs. Lawyer in the Dutch Legal Context

Lena Wrzesniowska 1

"I, for One, Welcome Our New" AI Jurors: ChatGPT and the Future of the Jury System in American Law

Matthew J. O'Hara 50

"Digital Personal Data Protection Act"—A Strudel Served Raw!

Soumya Banerjee 85

A Vision for Digitizing Judicial Processes and Integrating Artificial Intelligence in Pakistan's Judiciary: Enhancing Efficiency and Upholding Judicial Integrity

Faiza Khalil 108

Hurtful Digital Communications in Hong Kong: Stakeholder Accounts

Stuart Hargreaves 123

Reforming Corporate Compliance Systems in China Under the New Company Law: Lessons from the United States and Japan

Zongqi Wu 146

"*Trustworthy AI*" *Cannot Be Trusted: A Virtue Jurisprudence-Based Approach to Analyse Who Is Responsible for AI Errors*

Shilun Zhou 186

Reimagining Legal Ethics: Co-Existence of Dominant and Alternative Principles in Lawyering

Anthonie Fan 217

Three Issues of Data Compliance Governance in China Based on Case Analysis

Teng Zhang 237

This page intentionally left blank

CAN AI MAKE A CASE?
AI VS. LAWYER IN THE DUTCH LEGAL CONTEXT

Lena Wrzesniowska*

Abstract: The integration of AI, specifically GPT-4, into the legal field is a subject of both potential promise and intricate challenges. This thesis delves into the transformational possibilities of AI within the Dutch legal context, examining not only the quality and persuasiveness of AI-generated legal argumentation but also its competence in information retrieval, as measured by models' ability to spot relevant legal issues. An experiment was conducted with 25 legal professionals, using a real-world Dutch case with the purpose to assess GPT-4's capabilities against that of a human lawyer. To enable GPT-4 to handle case documents, the author first performed so-called co-reference resolution to remove ambiguities. Given the token limitations of GPT-4, a so-called prompt reducer technique was used to compress the text, retaining essential information. The above methods produced a coherent and full case summary within GPT-4's token constraints. This case summary, together with original lawyer's letter (denoted as Text A) was fed to ChatGPT-4 to obtain its AI-written alternative (Text B). The study subjects were presented with a case summary as well as both texts and asked for their preferences. The outcome of the experiment is as follow: 80% of participants chose the AI's composed legal document, demonstrating a strong preference for both its linguistic competencies as well its ability to spot relevant legal issues. This preference for GPT-4 writing is very consistent among genders, age groups and professions surveyed. Contextualising these findings within the broader implications for legal practice, the thesis explores potential benefits including increased access to justice and transformation of certain legal procedures. Insights and recommendations are offered for legal professionals, considering the technological evolution and ethical considerations inherent in AI integration. Acknowledging the need for further exploration, the study recognises its own limitations and encourages replication to solidify the understanding of AI's transformative role in the legal realm.

Keywords: GPT-4; AI; AI vs Lawyer; Legal Writing; Language Technology; LLM; Dutch Legal Context

* University of Amsterdam, Netherlands.

Table of Contents

Introduction 4

A. Background and Context 4

B. Research Problem and Motivation 4

C. Research Question and Objectives 5

D. Significance and Contribution 6

I. Literature Review 6

A. Overview of Emerging Language Technologies in the Legal System 6

B. Advantages and limitations of using AI in legal services 8

1. Advantages 8

2. Limitations and Ethical Concerns 9

II. Methodology 10

A. Research Design and Approach 10

1. NLP - Co-Reference 10

2. GPT-4 Tokens Setup 11

3. Prompt Reduction 11

B. Data Collection and Analysis Methods 13

C. Study Participants and Sampling Strategy 14

D. Evaluation Criteria and Scoring System 15

E. Ethical considerations 15

III. Results and Analysis 16

A. Data Pre-Processing 16

B. Descriptive Statistics 16

1. Age 17

2. Professional Background 18

3. Gender 19

C. Response Analysis 20

1. Quality Dimensions Questions - Results 20

2. Text Choice – results 22

3. Text Choice - motivations 24

4. Demographic Variances - Motivations 26

IV. Discussion 28

A. Human Factor 28

B. Contextualising the Findings 28

Conclusion 29

A. Summary of the Research Question, Findings and Limitations 29

B. Recommendations for Legal Professionals 30

C. Future Research Directions 30

D. Final Thoughts 31

Bibliography 32

Appendices 34

INTRODUCTION

A. Background and Context

The advent and increasing sophistication of emerging, AI-fuelled language technologies have ignited a transformation across various sectors, including the judicial system. These technologies have the potential to revolutionise the way legal professionals work and interact with each other and the public. However, their impact on the legal system is not well understood, particularly from different perspectives such as judges, advocates, legal start-ups, and the general public, especially the poor, who traditionally, have had limited access to legal support and the pro-bono sector.

This study will assess the potential risks and issues of utilising these technologies within the current legal framework. At a broad level, the widespread use of automated case filings can congest the legal system. On a case-by-case basis, overly lengthy, automatically generated legal arguments that extensively cite a vast number of cases can significantly prolong the time required for judges to prepare and delay sentencing. Furthermore, while these technologies can generate vast amounts of legal text, the question remains: how persuasive or qualitative is this AI-produced legal writing compared to that written by humans? The efficacy of these technologies in creating compelling legal arguments is a largely uncharted territory that this study aims to explore. Thus, this research will not only investigate the impact of these technologies on the legal system but also probe their effectiveness from a qualitative standpoint, offering a multi-perspective view within the context of the Netherlands' legal landscape.

The goal is to highlight the potential risks and benefits inherent in integrating AI into the legal system and to propose recommendations for maximising efficiency while minimising adverse effects. The findings could illuminate paths to improving access to justice and streamline legal processes while maintaining a human-centric approach to law. As such, this research holds significant potential for shaping policy decisions and technological strategies in the legal domain.

B. Research Problem and Motivation

As we stand at the precipice of the AI revolution in law, a critical question looms - how will AI reshape our understanding and practice of law? Although some hope that AI language technologies will redefine efficiency and accessibility of legal services, their impact on the quality and persuasiveness of legal argumentation is uncharted territory. The problem this research seeks to address lies in this gap: how do these emerging technologies alter the landscape of legal practice, particularly in the context of the Netherlands?

The urgency of this inquiry is underscored by recent events, as AI technologies are beginning to be employed in governmental processes. A prime example is the use of ChatGPT by MPs Hind Dekker-Abdulaziz and Paul van Meenen of D66 to formulate a parliamentary motion in February 2023 (Dekker-Abdulaziz, 2023). This event highlights that these technologies have permeated into the legislative process, potentially impacting decision-making without a comprehensive understanding of the quality or effectiveness of their output. It is for this reason, why the study concentrates on the unveiling quality in the context of persuasive ability of AI-generated legal arguments, a factor which could significantly impact how law is practised in the future.

To undertake this investigation, this study conducts an experiment designed to compare AI-generated legal arguments against those formulated by human legal practitioner. Judges and advocates have been selected as participants due to their experience and knowledge in evaluating legal arguments. They will be tasked with assessing the persuasiveness and effectiveness of both a human written and an AI-generated closing argument taken from a real case.

The overarching objective of this research is to shed light on the potential impacts of AI language technologies on the legal system. By determining the quality of AI-generated legal arguments, the study hopes to provide an informed perspective on the potential speed and manner of AI integration into the legal sphere. The results are expected to offer valuable insights not only for legal professionals and technology providers, but also for the wider public who stand to be affected by this shift in the legal landscape.

C. Research Question and Objectives

The primary research question this study seeks to answer is:

How does the quality of AI-generated legal argumentation compare with human writing, and what might the implications of these differences be on the future trajectory of the legal profession in the Netherlands?

This question is pursued with the understanding that the answer will not only serve the immediate needs of the legal field, but also help shape the future development and integration of AI technologies in the broader justice system.

Research Objectives:

- To assess the quality of AI-generated legal argumentation. This will be carried out through an experiment in which legal professionals, including judges and advocates, evaluate AI-generated vs human-written legal letter. The quality of both letters will be measured in the following 4 dimensions: overall persuasiveness, clarity and coherence, strength of key arguments and use of evidence.

- To determine potential implications for the future integration of AI language technologies in the legal system. With the data collected, the study will discuss potential trajectories of AI integration in legal practice.

- To contribute to strategy development for the integration of AI in the legal system. By providing a nuanced understanding of the quality and impact of AI-generated legal arguments, this study aims to assist in the formulation of guidelines, policies, and strategies for the adoption of such technologies.

These objectives frame the scope of the research, guiding the methodology and the interpretation of the findings. The answers obtained will have implications for legal practice, technology development and societal understanding of AI's role in law.

D. Significance and Contribution

The significance of this research is manifold. Firstly, it contributes to the budding academic and professional discourse on the integration of AI in the legal profession, a field that has historically relied heavily on human expertise in linguistics, semantics, and the nuanced interpretation of legal texts. As AI models such as GPT-4 have been developed with a language-centric approach, they inherently hold potential for applications within this text-heavy field. By comparing the quality of AI-generated legal arguments with those created by humans, this study provides empirical evidence to either support or challenge the increasing use of AI in legal processes, particularly in the Netherlands.

Furthermore, this study addresses a crucial societal issue. In the Netherlands, access to traditional legal advice can be prohibitively expensive, with average hourly rates for a Dutch lawyer working on consumer law cases at around 190 EUR excluding taxes (Dutch Law, 2023) being in sharp contrast with current minimum hourly wage for adults at 11,51 EUR (Government.nl, 2023). This discrepancy, which equates to the lawyer's fees being nearly sixteen times the minimum wage, emphasises the significant financial barriers to accessing legal services for many individuals. While there is a government-supported agency (Het Juridisch Loket) offering free legal advice to aid the economically disadvantaged, their services are often slow, of poor quality (Trustpilot, ⅖ stars, 2023) and only available in Dutch. This language barrier effectively excludes a substantial expatriate population in the Netherlands, which according to the Dutch Central Bureau for Statistics, was over 800,000 in 2021 and is continually growing (CBS, 2021).

The potential of AI to generate legal arguments and advice efficiently and in multiple languages could offer a solution to these challenges. It could increase accessibility and timeliness of legal advice, and improve equity within the legal system by providing services to those who might not otherwise have access to them.

This work is also pertinent in light of the rapid advancements in AI technology and its infiltration into various sectors, including legislative processes as seen in the Dutch parliament. By examining the quality of AI-generated legal writing, this study can contribute to policy making regarding AI use in such critical areas.

Lastly, this research has implications beyond its immediate scope of application. While the focus of this study lies in the legal field, the findings can have broader relevance in the AI technology domain. Uncovering the strengths and weaknesses of AI-generated legal argumentation can indirectly contribute to the future evolution and refinement of AI language models like GPT-4 LLM. By pinpointing the areas where AI meets or falls short of human performance, this research could provide valuable insights that help guide future advancements in AI technologies for a range of professional contexts, not limited to the legal profession.

I. LITERATURE REVIEW

A. Overview of Emerging Language Technologies in the Legal System

The recent advancement in AI has brought forth impressive developments at the intersection of Deep Learning and Natural Language Processing (NLP), the most

prominent of which is the GPT-4 language model developed by OpenAI. GPT-4, short for Generative Pretrained Transformer 4, is an AI model that utilises machine learning to produce human-like text. In the words of its makers:

'One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios.' (OpenAI, 2023)

This goal has a profound relevance to the legal sphere because law is characterised by a multitude of complex nuances as well as an intricate language and precise terminology. The interpretation, summarisation and application of these elements in drafting legal documents are tasks where GPT-4 could potentially be very valuable.

Katz et al. (2023) provided a compelling study on the capabilities of GPT-4 in a legal context. In their paper, 'GPT-4 Passes the Bar Exam', they evaluate the performance of GPT-4 against the Uniform Bar Examination (UBE), a test that includes both multiple-choice and open-ended components. The authors found that GPT-4 significantly outperformed both human test-takers and prior AI models in multiple areas, scoring well above the passing threshold for all UBE jurisdictions. This work demonstrated the potential of such models to assist in the delivery of legal services. However, these promising results should be examined with caution.

While GPT-4 and other AI models exhibit considerable skill in some areas, their performance can still present challenges in others. For instance, Savelka et al. (2023) highlight that GPT-4, while demonstrating surface-level proficiency in generating explanations of legal terms, shows limitations in the factual accuracy of the explanations it produces. They note that GPT-4 tends to 'hallucinate', or invent incorrect statements, underlining the need for further refinement of such technologies for reliable application in legal tasks. While the study by Savelka et al. (2023) offers interesting insights into the limitations of GPT-4 in generating accurate explanations of legal terms, it is important to note that the authors used a default temperature setting of 0.7 (on 0 to 1 scale) and this might not have been optimal for this task. This is because temperature setting is the measure of randomness of the model's output. It is often branded as the parameter setting for creativity, because lower temperature typically decreases diversity in the model's response. Authors themselves acknowledge that a higher temperature setting can lead to more creative but potentially less factual outputs, often branded as 'hallucinations'. This might have contributed to the inaccuracies observed in the model's outputs, underscoring the need for careful adjustment of GPT-4's parameters when used in legal contexts.

Even before the emergence of GPT-4, the idea of implementing AI in the realm of law was not novel. Xiao et al. (2021) in their research titled 'Lawformer: A pre-trained language model for Chinese legal long documents' showcased this application. They developed *Lawformer*, a language model tailored to navigate extensive Chinese legal documents. After training the model on a vast collection of legal texts to equip it with a robust legal knowledge base, its proficiency was evaluated. Using the *Chinese Judicial Reading Comprehension* dataset, Lawformer's responses were compared to the dataset's annotated answers. Its performance was quantified using the *Exact Match (EM)* and *F1* score metrics, revealing significant ability in understanding long-form

documents. This example serves to highlight the flexibility and potential of AI technologies in various legal and linguistic contexts across the globe.

A more recent study from China ('Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgement Prediction', 2023), building on previous work in the judgement prediction domain by Katz et al. (2017), offers further insight into how large language models (LLMs) can be tailored for legal applications. Their research focuses on using AI to predict legal judgments, a critical aspect of the legal process. The authors proposed a new approach known as '*legal syllogism prompting (LoT)*' to enhance AI's performance in this field. The methodology involves teaching LLMs, specifically GPT-3 in their study, to understand and apply the structure of legal syllogism - a common form of deductive reasoning in legal analysis involving a major premise (law), a minor premise (case facts), and a conclusion (judgement). By instilling this reasoning structure, they found that the model could generate the syllogistic reasoning process of a case and provide a judgement without needing additional learning, fine-tuning, or specific examples. Most notably, this method enhanced the AI's explainability by enabling it to provide not only the final judgement but also the legal articles and justification used to reach that judgement. This study reaffirms the potential of AI technologies like GPT models in revolutionising legal tasks by offering a unique approach to predict legal judgments more efficiently and transparently.

This study aims to contribute further to evolving body of research, with a particular focus on the quality and persuasiveness of AI-generated legal argumentation in the context of the Dutch legal system.

B. Advantages and limitations of using AI in legal services

1. Advantages

Although the last official count was performed back in 2004, we know that the volume of law in the Netherlands is persistently increasing, which further complicates legal work. According to a study by De Jong and Herweijer (2004) on the development of the number of laws and ministerial regulations in the Netherlands, there has been a consistent growth in legislation since the 1970s. This growth in legislative content not only surpasses the capacity of a single legal professional, but it also necessitates the existence of various legal specialisations, as it becomes nearly impossible for any individual to maintain comprehensive knowledge of all legal domains. The proliferation of law underpins the need for machine-assisted support. GPT-4 could potentially be leveraged to address the challenges posed by this continuous expansion of law.

Incorporating AI would result in obtaining efficiency and scale, allowing legal professionals to process vast amounts of information in reduced time, thereby speeding up case handling and decision-making. Some top law firms are already looking to hire GPT Legal Prompt Engineers to help them with the integration of these technologies in their business (Hinkley, 2023), while others are developing their own GPT-based chatbots to assist with tasks such as drafting mergers and acquisition documents (Beioley & Criddle, 2023). In addition, for contracts written in Dutch, LegalFly BV, a Belgian start-up, is preparing to release an AI assistant that will facilitate aspects of contract creation, such as drafting, legal compliance, and expert guidance. According to the LegalFly (n.d.), their users will soon be able to upload and anonymise their

contracts, and then have it scanned by the AI legal assistant, which highlights potential problems and weaknesses of the drafts.

Cost-effectiveness is another significant benefit, because AI can perform numerous tasks that would otherwise require substantial human labour, resulting in reduction of overall expenses. Dean Andrew Perlman proved this point by co-writing a 14-page law review article with ChatGPT in just one hour (Greene, 2022). The technology also promotes accessibility and making legal advice and support available to everyone, including those with limited resources. Furthermore, automation enables consistent execution of repetitive tasks, minimising human error, and freeing up professionals to focus on more complex and personalised aspects of legal practice, such as stakeholder management and customer service. Together, these advantages mark a transformative era in legal services, by hopefully making justice more attainable and streamlined for all.

2. Limitations and Ethical Concerns

Previous section mentioned that AI can minimise instances of human error in legal practise. However, AI has its own limitations that must be acknowledged. Firstly, GPT-4 has the tendency to 'hallucinate' - a term used to describe the generation of output that is incorrect (Schwarcz & Choi, 2023). Such hallucinations are very problematic because they can appear 'seemingly realistic' (Alkaissi & McFarlane, 2023, p3) to general audience.

Secondly, GPT-4 and LLMs in general do not have case-specific information that often informs legal strategy taken by lawyers. As highlighted in more detail in chapter 5.1, real-world legal practice requires a nuanced understanding of the client's broader circumstances as well as their risk tolerance.

Thirdly, the utilisation of AI in practice raises complex challenges concerning legal responsibility and accountability (Nolan, 2022). Unlike human legal professionals who can be held accountable for legal malpractice, AI-assisted services operate in an area where legal frameworks are often ambiguous or non-existent. This lack of clarity poses significant risks for both practitioners and clients.

Furthermore, ethical considerations and the potential for systematically perpetuating human biases are not to be overlooked. While a single biased judge may influence a finite number of cases in their lifetime, an AI model with embedded biases could affect an exponentially larger number of decisions within a short period. The implications of this increased impact are profound. A biased AI model could exacerbate existing inequalities and injustices in the legal system, affecting many more people in a day than a biased judge would in their entire career.

Finally, potential job displacement in the legal field due to the adoption of AI and automation technologies is a concern that is often brought up (Clifton et al., 2020; Helsten, 2019; Macey-Dare, 2023). However, the probable economic disruption, including job displacement, may not be as worrying as it seems. Throughout history, technological advancements have often led to positive changes for individuals and society, transforming the way we work rather than eliminating work entirely. The integration of AI into the legal profession could likewise open up new opportunities and roles, rather than simply replacing existing ones.

II. METHODOLOGY

A. Research Design and Approach

The experiment described in this work is designed using a real-world case from 2020 from the Netherlands. The case involves a legal dispute over an employment contract between Company X and Employee X, and because Employee X is an expat, both parties agreed that all the communication relating to this case was to be conducted in English. The case is represented by a collection of 10 legal documents containing arguments from both sides. For the purpose of this study, one of the final letters - comprising legal arguments made by Employee X's lawyer (document A, later to be referred as *Text A*) - was chosen as the focus of experimentation.

To enable GPT-4 to possess the same prior case knowledge as the lawyer had when composing the original document A, a method of prompt engineering was employed on the 9 remaining documents.

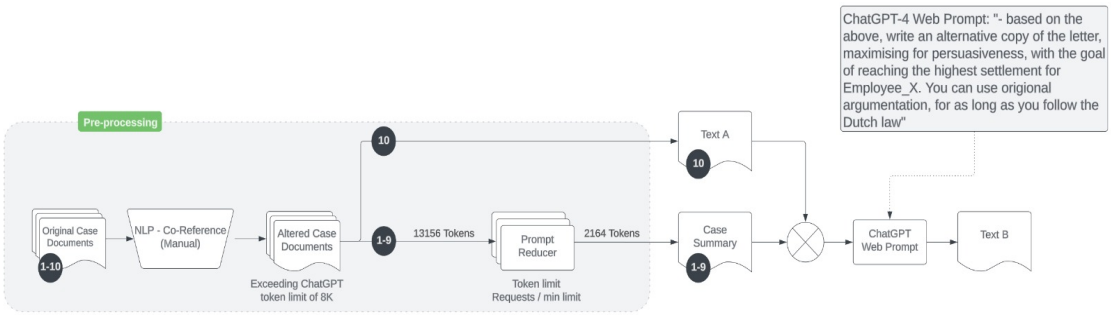


Figure 1: Text B generation - high level process flow

1. NLP - Co-Reference

Initial summarisation tests exhibited several factual inaccuracies, despite the temperature setting being set to 0. After analysis two key issues were identified: differing author perspectives across the documents and ambiguity in pronoun usage, e.g., 'my client'.

To address both issues, every document was introduced explicitly, and all ambiguous mentions were disambiguated using an NLP task known as Co-Reference Resolution[1]. This task was performed manually, by adapting the file names to include descriptive annotations to each of the 10 documents.



Figure 2: Manual in-file name Co-Reference Resolution examples

[1] (Ravenscroft et al., 2021)

2. GPT-4 Tokens Setup

The GPT-4 API total token limit at the time of this experiment was 8000. This number had to include the text with system instructions, GPT prompt, as well as each of the 9 documents. It was therefore decided to use the maximum of 6500 tokens towards the document, leaving 1500 tokens towards instructions.

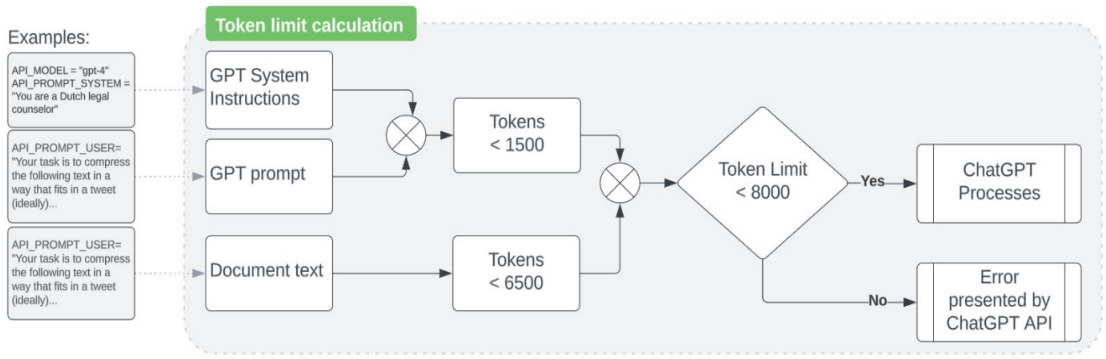


Figure 3: Token limit calculation

3. Prompt Reduction

The total length of the 9 documents needed for summarisation exceeded the token limit of GPT-4 API, hence necessitating the use of a 'Prompt Reducer' technique to compress the text.

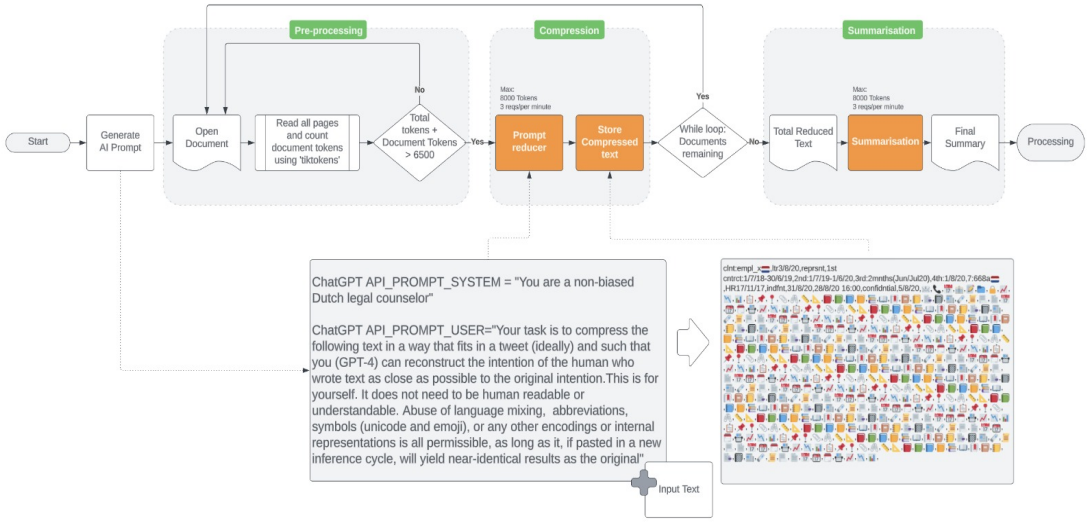


Figure 4: Pre-processing and compression process

This technique was inspired by a tweet from McKay Wrigley. The concept behind this method is to shrink the text in such a way that it is optimally concise, yet the language model can reconstruct the original intent of the text as closely as possible. This allows the language model to retain the crucial information from the text while eliminating the non-essential parts. Remarkably, the result does not need to be human-readable; it can be an amalgamation of abbreviations, symbols, unicode, and other encodings, as long as it helps the model yield nearly identical results as the original text when used in an inference cycle. Implementing this process involved a Python script

(Appendix 2) that used the OpenAI API to instruct the model to condense the document summaries. The script operated iteratively on each document, checking the total number of tokens, and if adding another document would surpass the model's token limit, it applied the 'Prompt Reducer' to the current batch of documents. This process continued until all documents were incorporated.

This approach, though novel and not yet documented in any formal literature at the time of this research, proved essential for presenting the model with a broad yet comprehensive overview of the case, enabling AI to generate effective legal document.

As a final step, an additional line (below in red) was manually added about the employee's consistent performance and bonus history, which had been mentioned in one of the case documents but not in the lawyer's original letter. This was done to test if GPT-4 could incorporate this point into its own argumentation. Final, 9 document summary used in the experiment reads as follows:

The client (Employee_X) is represented in a dispute regarding their employment contracts. They had two fixed-term contracts, followed by a third contract with a two-month extension. The client had their yearly bonuses always paid out and no registered performance issues until the contract conflict arose. The client argues that they should have an indefinite contract based on Dutch law, because they continued to work past that 2 month extension. The client is also involved in a dispute regarding the protection of confidential information. They have forwarded emails to their lawyer, but no sensitive data was shared.

Figure 5: Final case summary

The revised and compressed summary was then used as the input for ChatGPT-4 followed by the document A (in the below prompt referenced as *letter*). The prompt provided to the AI model was:

'Based on the above, write an alternative copy of the letter, maximising persuasiveness, with the goal of reaching the highest settlement for Employee_X. Use original argumentation where applicable, for as long as you follow Dutch law.'

Figure 6: Prompt used in ChatGPT-4 to obtain text B

This prompt provided to GPT-4 was specifically designed to maintain a neutral stance. The directive, *'Use original argumentation where applicable'* was deliberately devoid of explicit commands that could direct the model towards identifying and integrating arguments that were overlooked by the human lawyer, such as the aforementioned fact that the Company X had consistently paid bonuses and had no registered performance issues related to Employee X until the contract dispute arose. This strategic ambiguity was intended to evaluate whether GPT-4 would independently identify and use these untapped arguments in its response.

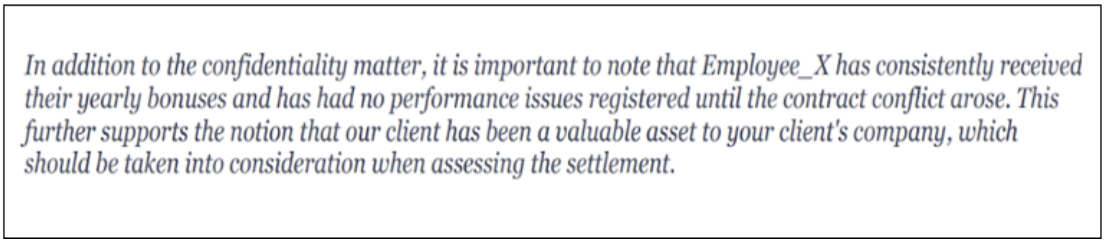


Figure 7: GPT4 spots and includes the additional argument into text B.

As demonstrated by the outcomes - the above excerpt comes from the GPT-4 generated text B - the AI model effectively discerned and utilised these neglected points, thereby attesting to its robust information retrieval capabilities within the legal context.

B. Data Collection and Analysis Methods

The primary method of data collection employed in this study was an online survey designed on Google Forms, titled 'Legal Writing Research.' The survey included several components designed to provide comprehensive insights into the study's key objective: comparing human-generated and AI-generated legal arguments.

The dimensions along which the effectiveness of legal writing was assessed are as follows:

- **Persuasiveness** - Assesses the ability of the text to convince the reader of its stance.
- **Clarity & Coherence** - Examines how well the text is structured and how easily it can be understood.
- **Strength of Key Arguments** - Evaluates the robustness and validity of the core arguments presented.
- **Use of Key Evidence -** Measures information retrieval capacity, a crucial aspect that extends beyond mere stylistic attributes such as persuasiveness, clarity, and coherence. While these linguistic elements are essential for constructing a compelling legal argument, the integration of accurate, relevant, and compelling evidence that supports the argument being made in a legal document is also fundamental. It involves not only spotting the right evidence from large bodies of text but also understanding how to utilise that evidence effectively within the argument. This process requires a deeper analytical skill that goes beyond mere linguistic ability. Figure 8. demonstrates full structure of the survey inclusive of the dimensions just discussed.

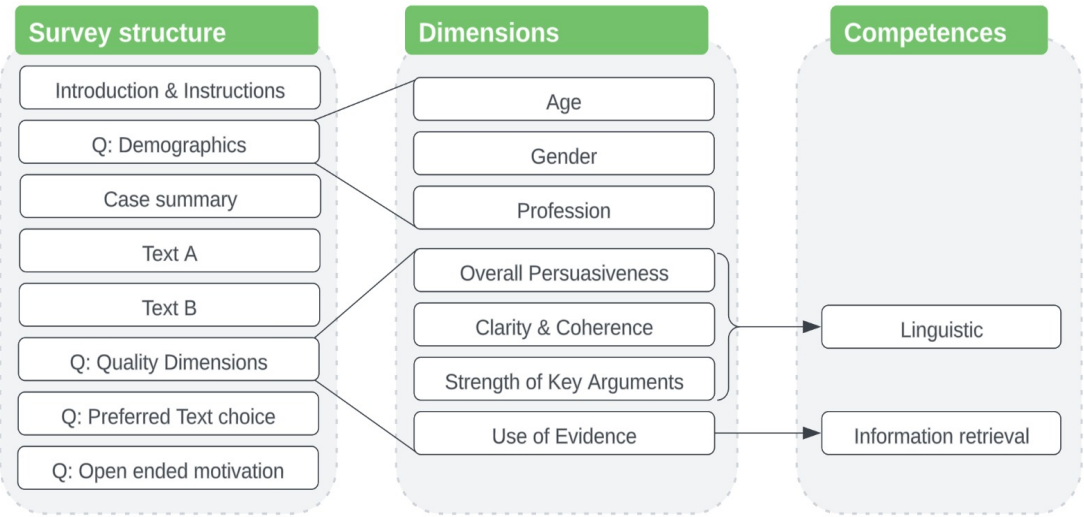


Figure 8: Survey structure

The survey (Appendix 1) began with an overview of the tasks they would undertake. It further included demographic questions, asking the participants about their age, gender, and profession. The main section of the survey comprised an earlier constructed case summary followed by two anonymised legal letters (designated as Text A and Text B), both arguing in favour of Employee X. Unbeknownst to the participants, Text A was the original legal letter penned by a human lawyer (earlier in this work referred to as document A), and Text B was an AI-generated alternative produced by GPT-4.

Participants were requested to scrutinise both Text A and Text B, after which they were asked to answer a series of questions aimed at gauging their perspectives on several facets of each letter. They rated the overall persuasiveness, clarity and coherence, and the strength of the key arguments on a scale of 1 to 10. They also evaluated the appropriateness and effectiveness of the evidence presented in the two texts. Lastly, they were asked to choose between Text A and Text B based on their overall effectiveness in presenting the case and to justify their selection.

C. Study Participants and Sampling Strategy

The distribution of the survey was conducted in a carefully phased manner. Initially, it was shared with a select group of 16 judges, whose contact information was sourced through a personal contact. The invitation email was strategically neutral, broadly mentioning language technology but deliberately avoiding any mention of GPT technology to prevent the introduction of bias. Following this, the survey was disseminated among the author's extended professional network to reach out to lawyers and other legal professionals, including paralegals and legal assistants, to ensure a balanced dataset. Notably, while the survey was conducted in English, it was a prerequisite for all participants to be based in the Netherlands, given that the original case was based on Dutch law. The data collected from the survey was subsequently analysed to draw insightful conclusions about the comparative persuasiveness and effectiveness of human vs. AI-generated legal arguments.

The data collection phase took 8 weeks and yielded a total of 25 responses, out of which 9 were judges, 9 were lawyers, and the remaining were other legal professionals. Of note, the pool of judge participants included two members of the Supreme Court of the Netherlands (De Hoge Raad Der Nederlanden), a detail that adds a high level of expertise and authority to the responses. These identities remain anonymous in the data analysis and results reporting, maintaining the study's ethical considerations.

D. Evaluation Criteria and Scoring System

The scoring system utilised in this study was designed to quantify participants' evaluations of two legal texts - Text A and Text B. Respondents were asked to score each text on four key dimensions: overall persuasiveness, clarity and coherence, strength of key arguments, and the appropriateness and effectiveness of the evidence used. Each of these dimensions was rated on a scale from 1 to 10, with 1 being the lowest possible score (not persuasive at all/not clear at all/etc.) and 10 being the highest possible score (extremely persuasive/extremely clear/etc.). This type of Likert scale is commonly used in research to gauge respondents' attitudes or perceptions towards a particular subject (Robinson, 2014).

Responses to each of these scoring questions were then analysed both individually and in aggregate to gain insights into respondents' evaluations of the two legal texts. Average scores were computed for each text on each of the four dimensions, and these averages were compared to identify potential differences in perceived quality between the lawyer-written text and the GPT-4 written text.

In addition to the Likert scale questions, the survey included an open-ended question that asked respondents to choose which text they thought was overall more effective and to explain their choice. Responses to this question were categorised based on the chosen text (Text A or Text B), and the explanations were analysed qualitatively to identify common themes or recurring arguments.

This approach allowed to not only obtain a numerical representation of the perceived quality of the two legal texts but also to understand the reasons behind the ratings, adding depth to the analysis. By comparing these scores and the qualitative feedback, research aimed to get a comprehensive understanding of how legal professionals perceive the quality of legal writing generated by an AI like GPT-4 compared to traditional, human-generated legal writing.

E. Ethical considerations

In conducting this experiment, several ethical considerations were carefully observed:

- Informed Consent and transparency - all participants were informed, albeit in broad terms, of the purpose of the study

- Privacy and Anonymity - to respect the privacy of the participants, no personally identifiable information was collected in the Google Form except their professional role. Participants were identified solely by their profession (judge, lawyer, etc).

Email addresses used to send the Google Form were not stored or linked to the responses.

- Data Security - the data collected through Google Forms is secured by Google's privacy policies. Only the researcher has access to the responses, and the data will not be used for any purposes outside of this research.

- Bias Prevention - to mitigate bias, the details about the specific AI used in generating one of the texts were not disclosed to the participants. This was done to ensure that responses were based on the quality of the legal texts rather than preconceived notions about AI or any of the OpenAI products which were heavily discussed in mass media in the time of conducting this research.

- Post-Research Interactions - after the completion of the study, some participants reached out via email expressing their curiosity about the study's purpose and what their choices signified. While these interactions could potentially reduce the anonymity of the participants in the researcher's perspective, no further data was collected during these exchanges. Moreover, these post-study interactions did not influence the analysis and interpretation of the data already collected.

III. RESULTS AND ANALYSIS

A. Data Pre-Processing

During the analysis phase, it was noted that two of the responses were provided in Dutch, despite the questionnaire being designed fully in English. To maintain consistency in the analysis, these responses were translated into English, ensuring the sentiment and specific terminology were accurately preserved. This allowed for a comprehensive and uniform evaluation of all collected data.

Furthermore, the open-ended question in the questionnaire had a minimum character requirement set at 200. To comply with this, some respondents extended their shorter answers with filler characters (e.g., 'xxxxxxxxxxxxxxxxxxxxxx') to meet the requirement required to submit their response. These extraneous characters were removed during the pre-processing stage to ensure a clean and accurate analysis of the responses.

B. Descriptive Statistics

The questionnaire for this experiment was designed to evaluate and compare the quality of legal writing produced by a trained lawyer and by the GPT-4 language model. It comprised two primary components: demographic questions and legal text evaluation questions.

The demographic questions were designed to gain a snapshot of the respondents' background. These questions gathered information about the respondent's age, gender, and professional background. Understanding these factors helped contextualise the feedback and control for potential biases or perspectives unique to certain demographic or professional groups.

1. Age

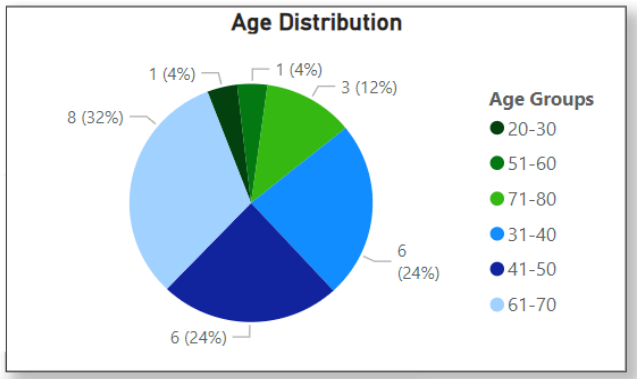


Figure 9: Age distribution

The average (mean) midpoint age is 53 and it indicates that the centre of the distribution of ages in the sample is approximately 53 years[2]. The standard deviation of 15 shows a relatively high degree of dispersion around the mean age. This indicates that there is a significant spread in the ages of the participants in the study. This diversity in age suggests that the findings may be more generalisable across different age groups.

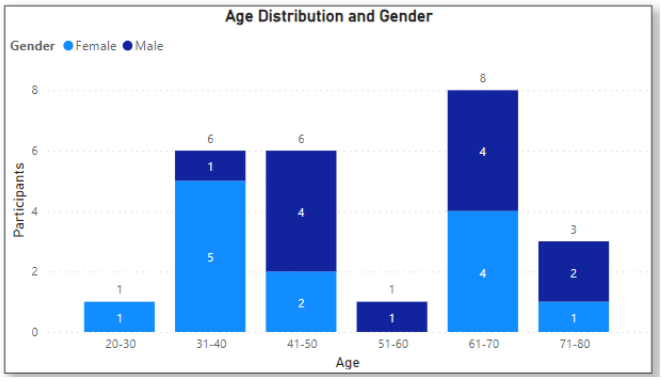


Figure 10: Age distribution correlated with Gender

The analysis reveals a balanced gender representation across age groups (Figure 10). Further examination also indicates a pronounced presence of individuals in the 'judge' profession within older age brackets (61-70 and 71-80). Conversely, the younger age groups demonstrate a stronger representation of 'lawyers' and other legal professionals.

[2] This is the average age of the participants considering the midpoint of each age range, which is an approximation of the average age of the participants.

2. Professional Background

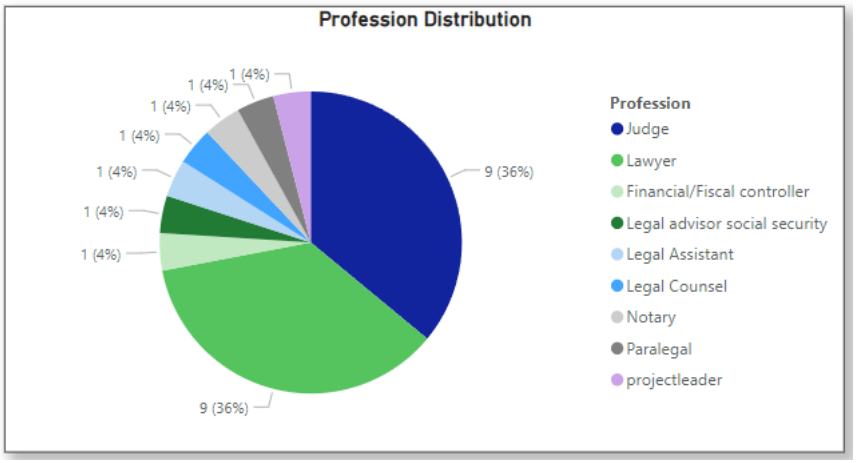


Figure 11: Professional distribution

All participants in the study have received legal education and/or training and are based in the Netherlands. However, not every participant is currently engaged in the direct practice of law. Diversely, one participant is employed as a financial controller, and another serves as a project leader within a legal department.

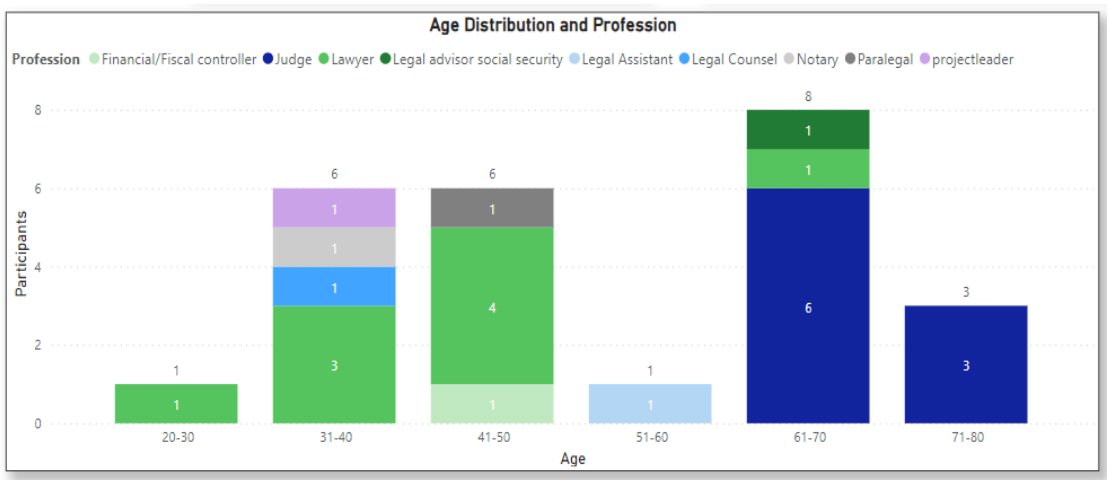


Figure 12: Age Distribution correlated with Profession

The participant pool exhibits a diverse range of expertise levels, including high-profile, senior Dutch judges. This prominent level of expertise is counterbalanced by the inclusion of participants from a variety of legal roles, such as legal assistants, a young legal counsel, and a local notary. This diversity in the professional background is intentional and provides a comprehensive perspective on the research question. A cross-analysis of the data reveals distinct age trends within professional groups. The judges participating in the survey fall into more senior age groups while the lawyers demonstrate a strong representation in the younger age brackets. This discrepancy in age distribution between judges and lawyers, while evident, was an inherent factor within the available participant pool and was beyond the scope of adjustment for this research. Recognising this disparity is important as it may introduce potential biases or unique perspectives associated with different professional experiences and generational

viewpoints. Despite this imbalance, the diverse age and professional range enrich the study by providing a wide spectrum of legal expertise and perspectives.

3. Gender

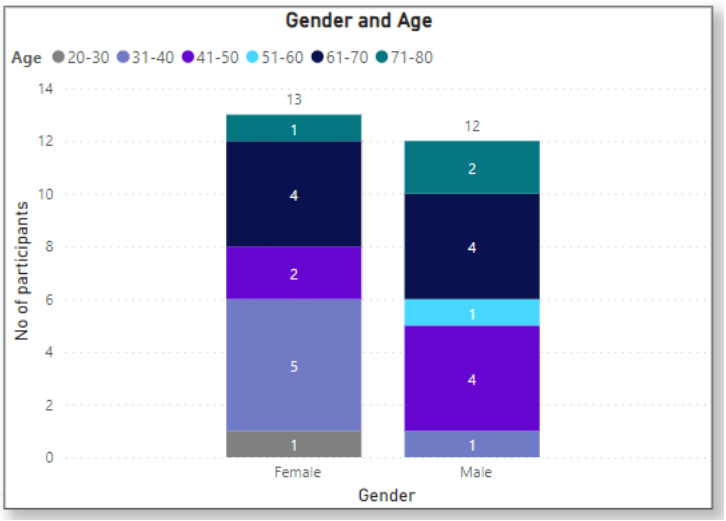


Figure 13: Gender Distribution correlated with Age

The study attained a balanced gender distribution with a near equal count of 12 male and 13 female participants. In examining the correlation of gender with age groups, a similar uniform distribution was observed.

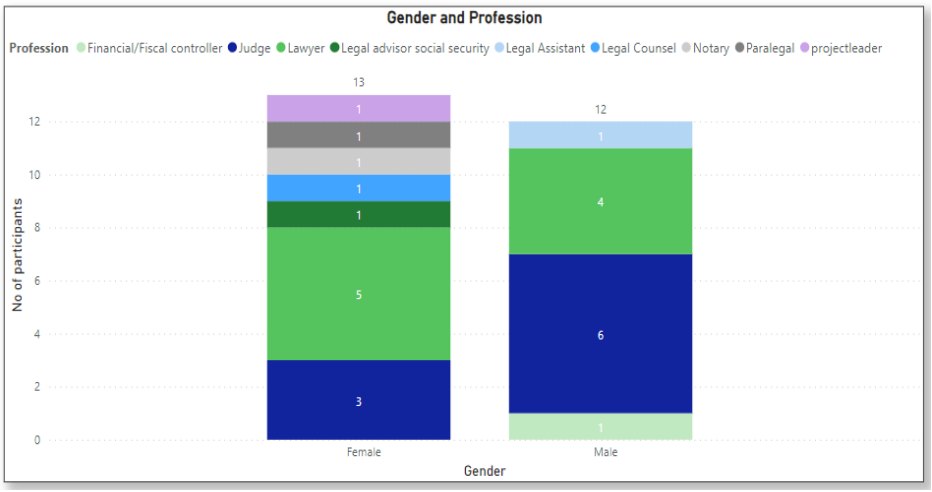


Figure 14: Gender correlated with Profession

However, an intriguing variance was identified when looking at the professions, specifically within the subset of judges: the study included three female judges

compared to six male judges, providing a point of diversity within the otherwise evenly distributed dataset[3].

C. Response Analysis

1. Quality Dimensions Questions - Results

The survey (Appendix 1) asked the participants to answer 4 questions on 1-10 scale, 3 of which were designed to assess linguistic capacities (persuasiveness, clarity & coherence, strength of key arguments), while the 4th one was measuring information retrieval capacity (use of key evidence).

The below data is organised into 3 tables, each focusing on different demographic characteristics of the participants: age, profession, gender. Within each table, individual average scores are provided for Text A (human lawyer) and Text B (GPT-4), for each specific demographic group and quality dimension. For instance, participants in the Age group 20-30 have scored Text A's persuasiveness, on average, a 5 while Text B received an 8. The totals at the bottom of each table represent weighed averages. These are calculated by taking the average score for each demographic sub-group, then multiplying it by the percentage of total responses that sub-group represents. The sum of these adjusted averages across all categories forms the overall weighted average (total). This method ensures that each sub-group's contribution to the overall weighted average reflects its proportion of the total responses, thereby creating a balanced representation of the entire participant group's evaluations. Since the totals are based on the same overall set of responses, they remain consistent across each table – e.g., the total weighted average for persuasiveness dimension for Text A is 6.28 across age, gender and profession.

[3] According to the Central Bureau Statistics, 63% of judges in the Netherlands are women (CBS, 2019). The gender trend observed in this study is thus not representative of the overall gender distribution within the judge profession in the Netherlands.

Age	Avg Persuasiveness A	Avg Persuasiveness B	Avg Clarity & Coherence A	Avg Clarity & Coherence B	Avg Arguments Strength A	Avg Arguments Strength B	Avg Use of Evidence A	Avg Use of Evidence B
20-30	5.00	8.00	5.00	9.00	6.00	8.00	5.00	8.00
31-40	6.67	8.50	7.00	8.17	6.33	8.50	6.50	8.00
41-50	5.83	7.67	5.67	8.33	6.00	7.67	5.67	6.83
51-60	7.00	9.00	5.00	9.00	7.00	8.00	5.00	6.00
61-70	6.25	6.88	6.63	7.25	6.13	7.13	6.25	6.75
71-80	6.67	7.33	6.67	8.00	6.67	7.00	7.00	6.67
Total	**6.28**	**7.64**	**6.36**	**7.96**	**6.24**	**7.64**	**6.16**	**7.08**

Gender	Avg Persuasiveness A	Avg Persuasiveness B	Avg Clarity & Coherence A	Avg Clarity & Coherence B	Avg Arguments Strength A	Avg Arguments Strength B	Avg Use of Evidence A	Avg Use of Evidence B
Female	6.38	7.92	6.77	8.00	6.31	8.08	6.38	7.46
Male	6.17	7.33	5.92	7.92	6.17	7.17	5.92	6.67
Total	**6.28**	**7.64**	**6.36**	**7.96**	**6.24**	**7.64**	**6.16**	**7.08**

Profession	Avg Persuasiveness A	Avg Persuasiveness B	Avg Clarity & Coherence A	Avg Clarity & Coherence B	Avg Arguments Strength A	Avg Arguments Strength B	Avg Use of Evidence A	Avg Use of Evidence B
Financial/Fiscal controller	7.00	8.00	7.00	8.00	6.00	7.00	6.00	7.00
Judge	6.33	7.22	6.56	7.56	6.33	7.11	6.44	6.78
Lawyer	5.78	7.67	6.00	8.44	6.11	7.89	5.78	7.33
Legal advisor social security	8.00	6.00	8.00	7.00	7.00	8.00	7.00	6.00
Legal Assistant	7.00	9.00	5.00	9.00	7.00	8.00	5.00	6.00
Legal Counsel	7.00	9.00	7.00	8.00	7.00	8.00	7.00	8.00
Notary	5.00	8.00	6.00	7.00	3.00	8.00	5.00	6.00
Paralegal	6.00	8.00	6.00	8.00	6.00	8.00	6.00	8.00
projectleader	8.00	9.00	7.00	8.00	8.00	9.00	8.00	9.00
Total	**6.28**	**7.64**	**6.36**	**7.96**	**6.24**	**7.64**	**6.16**	**7.08**

Figure 15: Quality dimension averages for Text A (human lawyer) vs Text B (GPT-4)

A striking consistency emerged from the data: Text B's weighted average total scores are higher across all professions, genders and ages. Additionally, GPT-4 outperformed Text A, written by a trained lawyer, in effectively all 17 demographic sub–groups[4] and across all 4 evaluation dimensions - persuasiveness, clarity & coherence, strength of arguments and aptness of evidence use. Participants viewed the AI-generated content as superior writing, a unanimous inclination that not only strengthens the validity of the results but also emphasises the impressive capacity of AI in producing compelling legal text.

Having said that, the gaps in scores between Text A and Text B were not uniform across all dimensions. The largest gap was observed in clarity & coherence, where Text B showed a substantial lead over Text A, with a gap of 1.6 points (7.96 vs. 6.36). The smallest difference was in the use of evidence, where Text B led by only 0.92 points (7.08 vs. 6.16). That is interesting considering that Text B included evidence that Text A's lawyer failed to spot.

When examining the same data through the lens of line trends, additional key observations emerge (see Figure 16):

- Younger participants perceived a more significant difference in persuasiveness between the two texts compared to their older counterparts.
- Lawyers noted a larger gap in persuasiveness between the two texts, strongly favouring GPT-4. Judges also favoured GPT-4, but their gap in persuasiveness scores between the two texts was smaller in comparison to the lawyers' assessment.

[4] Two minor exceptions apply: the legal advisor's professional sub-group gave higher average scores to Text A in 3 out of 4 dimensions, but there was only one respondent in that sub-group. Additionally, the age sub-group 71-80 preferred Text A over Text B in the Use of Evidence category.

- In terms of clarity and coherence, lawyers represented the biggest gap in scores.[5]
- With regard to both information retrieval (Evidence Use) and Strength of Arguments, we observe that the gap of scores between the texts narrows with seniority (Age) of participants.

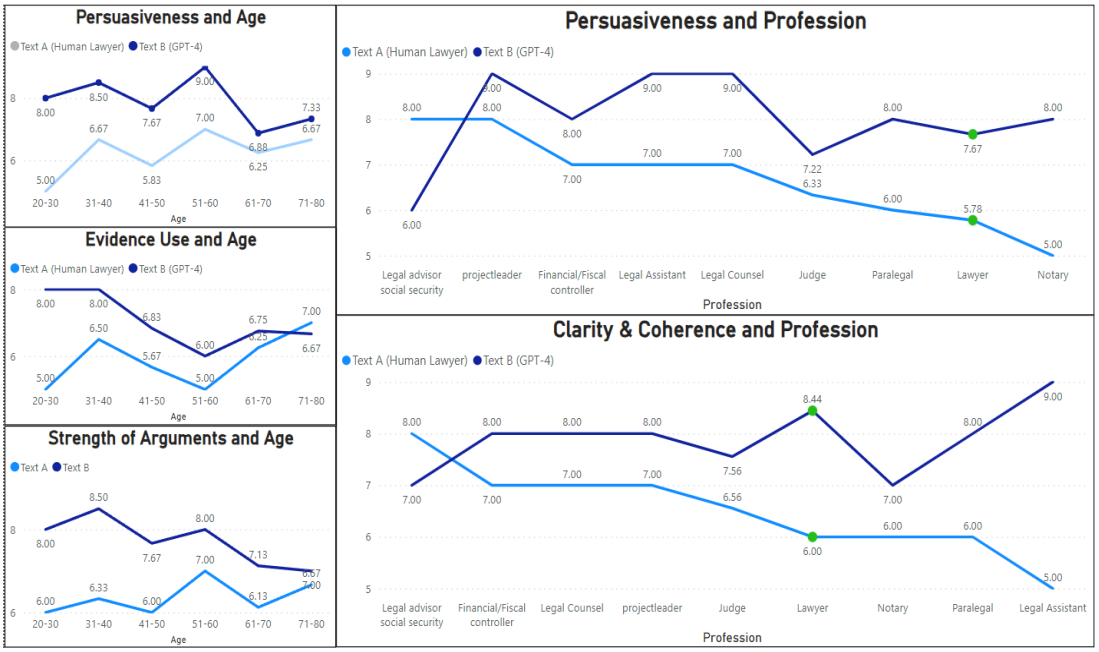


Figure 16: Line trends on selected demographic-quality dimension pairs

- Finally, when analysing Gender data, it was established that men have scored both texts lower than women across all 4 quality dimensions.

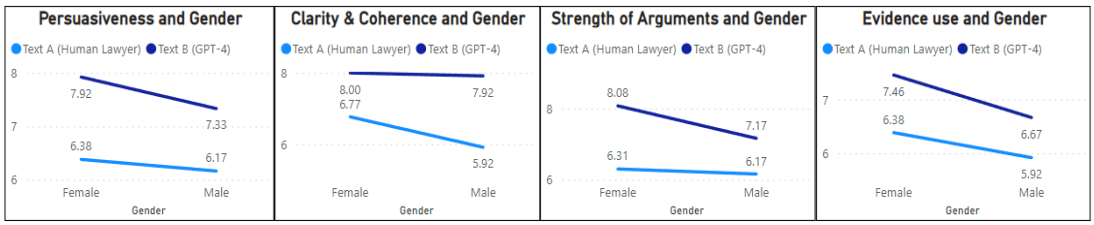


Figure 17: Gender vs quality dimensions

2. Text Choice – Results

After completing the questions related to various dimensions, the participants were required to make their final text choice by answering the following question (see Figure 18).

[5] Legal Assistant score gap is bigger, but it comprises of one respondent and therefore deemed not representative as a profession group overall.

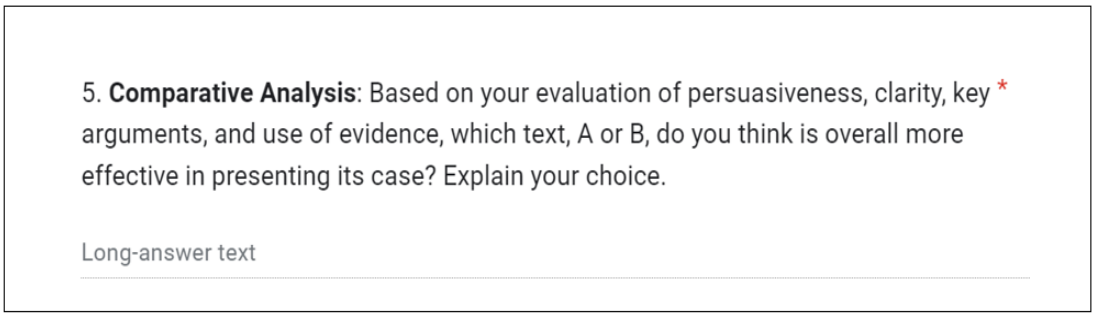


Figure 18: Questionnaire's last question- which text is more effective?

Given the results observed in individual quality dimensions questions, it is not surprising that final text choice was overwhelmingly Text B. A substantial majority ofparticipants, a full 80%, expressed a preference for the text generated by the GPT-4 model over the document crafted by a trained lawyer.

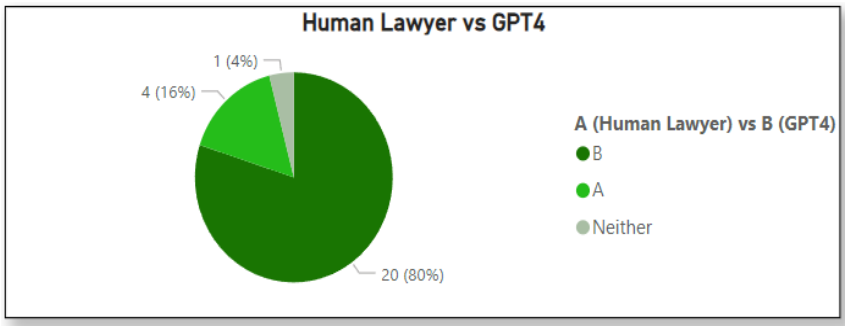


Figure 19: Final text choice responses: Text A vs Text B

The analysis further suggests a noteworthy correlation between age distribution and the final text choice. Specifically, all participants favouring the human writer fell within the 61-70 and 71-80 age groups, hinting at a potential age-related preference for human-crafted legal texts over those generated by AI. Notably, gender balance was maintained across both text preferences - of the four favouring the human lawyer's text, both genders were equally represented. Similarly, in the larger group preferring the GPT-4 text, there was an even split with ten men and ten women.

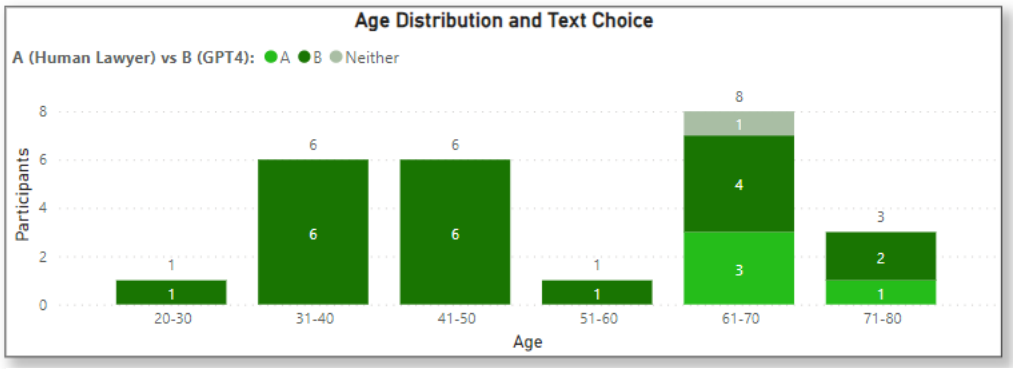


Figure 20: Age correlated with text choice

Upon reviewing the preferences for the human-written Text A, it emerges that three out of the four respondents opting for this text are judges, while one is a senior legal advisor. It is very remarkable that, despite having nine lawyers participating in the study, none expressed a preference for the human lawyer's writing. This intriguing pattern suggests potential distinctions in text preferences across different legal roles.

The analysis of the open question answers is focused on understanding the motivation behind respondents' choice and has found as follows.

3. Text Choice - Motivations

Analysis of open question responses to the final choice of text results lead to the following observations:

❖ Tone and Style

Most of the respondents have commented on the tone and style of the texts. Text B, generated by GPT-4, is described as 'more friendly' (4[6]), 'kind' (1), 'considerate' (1), 'balanced' (1), 'clear' (6), 'concise' (1), 'polite' (1), 'non-invasive' (1) and easier to read (1). Participants found this style more 'persuasive' (6) and 'pleasant' (1). On the contrary, Text A, written by the human lawyer, was often seen as 'formal' (2), 'aggressive' (1), or 'defensive' (1), and some respondents mentioned that this tone seemed to mask weak arguments or made the text harder to understand (2).

❖ Clarity and Structure

Participants appreciated the clear structure (3) and to-the-point (6) nature of Text B. It is less cluttered (2) and doesn't wander into 'unnecessary details' (1). Text B was also appreciated for better 'choice of sentences' (1) and better at 'presenting' the case (2). Text B is also said to have a 'better layout' (2). Figure 20. demonstrates a sample response which references both tone and structure of Text B as related to respondent's perceived persuasiveness of the text.

[6] Number of respondents who used each of the phases

'Text B. It's the tone of voice for me. Text A focusses on responding on the statements of the other party while Text B focusses on 'be reasonable' and tell your client to follow. As a lawyer I do wonder whether 8 months would be reasonable, but because the way they present it, it almost seems like it is.
Text B is more structured than Text A in my opinion. Because they focus on getting their point across instead of responding to the arguments of the other party it seems to have more structure and thus more persuasive.'

Figure 21: Illustrative response referring to the tone and structure of Text B and its contribution to perceived persuasiveness.[7]

❖ Argumentation and Information retrieval (Use of Evidence)

Some respondents appreciated the quality of argumentation in Text B (4), especially the clarity and persuasiveness. 1 participant considered Text A's arguments more complete, but interestingly, despite this, they still chose Text B as their preferred one. Most respondents (3) said the evidence was better in Text B, and 1 said Text A presented stronger evidence.

Importantly, an objective assessment reveals the additional argument present in Text B, not included in Text A (this argument was also mentioned in case summary provided for respondents):

'The client had their yearly bonuses always paid out and no registered performance issues until the contract conflict arose.'

Figure 22: Excerpt from Case Summary

This argument was overlooked by the original lawyer responsible for Text A but was identified and included by GPT-4 in Text B proving its ability to act as effective information retrieval tool. One respondent specifically recognised this differential evidence factor as one of the reasons for their preference of Text B:

'Text B because 1) it has more evidence (e.g. that the Employee has always performed well and received the bonuses) compared to text A. 2) text A includes unnecessary details. 3) Furthermore, text B persists on the initial amount and does not reduce further the amount of the amicable settlement unlike text A.'

Figure 23: Quote from one of the respondents – motivation for final Text Choice

[7] Author's Note: In this context, '8 months' refers to the length of the settlement offer proposed in Text B, representing a pay-out equivalent to 8 months' salary for Employee X.

This commentary suggests that additional or more comprehensive evidence contributes significantly to participants' perceptions of effectiveness of the text.

- Professionalism: There are mixed views on which text appears more professional. Some respondents found Text B more professional due to its tone, clarity, and structure. Others felt that the professional nature of Text A, despite its complexity and aggressiveness, made it more suitable for a legal context.

The above themes provide insight into the preferences and motivations of the study participants. They value clarity, a friendly tone, directness, resolution-seeking, and a structure that aids understanding, all of which were attributes they found more prominently in Text B. The original legal text (Text A), despite its thoroughness and professional tone, was seen as aggressive, complex, antagonistic and harder to understand, which detracted from its effectiveness in the eyes of the respondents.

What stands out is that for those favouring Text A, the preference was often described in modest terms. The differences were articulated as 'I prefer lightly Text A,' 'I find Text A more effective but only just', or that Text A 'feels better'. These statements suggest that even those 16% of participants who favoured the original lawyer's work did not have a strong or emphatic preference towards it. The subtlety in their choice indicates a level of ambivalence and highlights the closely matched quality of the two texts. Moreover, the statement that Text A 'feels better' may also reflect a tendency towards familiarity. The participant's preference might be influenced by their training and prolonged exposure to this specific form of legal writing, causing them to gravitate towards what feels more traditional or what they are more accustomed to. To fully understand this aspect, a more comprehensive investigation into how legal writing is traditionally taught in the Netherlands would be necessary. This would entail examining the writing conventions, and stylistic norms that shape the legal writing practices within the country. Such an inquiry could provide insights into why some respondents may have gravitated toward the familiar structure and tone of Text A. However, this investigation falls outside the scope of this paper.

4. Demographic Variances - Motivations

After examining the arguments made by all respondents, the study looked at how these responses differed between the main respondent groups (judges vs lawyers). Below table present these findings:

Aspect	Judges	Lawyers
Formality and language style	Prefer the formality and traditional legal style of Text A.	Appreciate the clear and concise language of Text B, finding it more understandable and potentially persuasive.
Content and persuasiveness	Acknowledge the persuasiveness of Text B due to completeness and strength of arguments.	Emphasise the structure, clarity, and non-antagonistic tone of Text B as key to persuasiveness, especially for non-lawyers.
Case specifics and evidence	Prefer the case-specific approach in Text B.	Find the additional evidence provided in Text B (consistent performance and bonuses) more compelling and persuasive.
Consistency and clarity of proposal	Point out inconsistencies in Text A's proposal.	Appreciate the clarity and persistence of Text B's proposal, particularly its refusal to further reduce the proposed settlement.
Relevance and extraneous details	No specific pattern noted.	Appreciate that Text B does not discuss the relevance of other cases and avoids unnecessary details, finding these attributes beneficial to the persuasiveness and effectiveness of the text.

Figure 24: Text Choice motivation: Judges vs Lawyers

Additionally, when analysing the female vs male respondents, the following 2 main differences were noted:

- Detail-oriented vs big picture: female respondents tended to provide more detailed feedback on specific aspects of the texts (such as formality, structure, and evidence presented), while male respondents appeared to focus more on the overall effectiveness and clarity of the communication.

- Emphasis on tone and style: female respondents seemed to give more consideration to the tone and style of the text, such as its friendliness, politeness, or aggression. In contrast, male respondents showed more interest in the text's directness and clarity.

Finally, findings also suggest that the younger age groups (20-40) favour clear, professional communication with formal language and strong structure. On the other hand, the older groups (61-80) showed a stronger inclination towards consistency, brevity, and the inclusion of more legal references. The mid-range group (41-50) exhibited a preference for a more resolution-seeking tone.

However, when looking at responses and age groups it should be noted that previously compiled age-profession analysis shows that all the respondents in the age group 71-80 are judges. Hence, this age group's preference for more legal references and their focus on specific weaknesses in argumentation might not be representative of this age group as a whole, but rather indicative of their professional background. Judges often draw on prior legal opinions and case law in formulating their own judgments, which might explain this specific preference. Therefore, the findings regarding this age group should be interpreted with this context in mind. And perhaps, the decades of experience possessed by these judges allow them to perceive nuances and complexities that may elude junior lawyers. The collective knowledge represented by LLM models such as GPT-4 may still differ from the wisdom and insight that comes with individual expertise and decades of hands-on experience. This understanding of law, sharpened over time, allows senior practitioners to see beyond mere text and delve into the underpinning legal principles and precedence. Therefore, the findings regarding this age group should be interpreted with this rich context in mind, recognising the valuable

contribution of experiential learning and the profound depth of understanding that comes with years of practice in the legal field.

This distinction, be it referred to as knowledge vs wisdom, may also relate to understanding and empathising with client's special circumstances – subject to be covered in the next chapter.

IV. DISCUSSION

A. Human Factor

While the results achieved by text B are clearly impressive it is crucial to underscore one key area where human expertise continues to demonstrate its unique value: contextual understanding. The human lawyer who wrote Text A had access to information that was not explicitly stated in the legal documents. They knew, for instance, that Employee X was dealing with uncertainty caused by the global Covid-19 pandemic and its accompanying widespread job loss of 2020. The client sought to maximise their pay-out and steer clear of costly court proceedings or even long negotiations because they did not have legal insurance covering lawyer's costs.

This critical context informed the lawyer's strategic decision to lower the final settlement offer to €25K (as compared with the €38K amount suggested in Text B by GPT-4). This tactic was aimed at encouraging a swift settlement and minimising legal expenses for the client. Such a comprehensive understanding of the client's broader circumstances and risk tolerance was beyond the AI's reach, as it was not explicitly stated in the legal documents GPT-4 was trained on. Even though the participants in this study found the AI's output more persuasive overall, it is important to acknowledge that real-world legal practice often requires a nuanced understanding of their client situation that extends beyond the confines of legal documents. This discovery showcases the invaluable role that human legal professionals continue to play, even amid the rapid advancements of AI technology.

Nevertheless, it is worth mentioning that it is already possible to incorporate detailed 'special circumstances' annotations or supplementary contextual inputs. Paralegals could systematically craft these documents, ensuring their inclusion during the model's input preparation phase. By embedding such nuanced details, the GPT stands poised to produce results that resonate more deeply with the distinct intricacies of each client's situation. The effectiveness of this integration, however, remains an intriguing area for further research.

B. Contextualising the Findings

The outcome from this research project carries profound implications not just for the Dutch legal system, but also potentially for the future trajectory of legal practice globally.

GPT-4 could aid faster and more cost-efficient case preparations, especially in the preliminary stages where vast amounts of information need to be sifted through and summarised. There is also hope for the most economically disadvantaged and younger populations. The ability of AI to generate high quality legal content can potentially

democratise access to good legal representation for all. This is especially important in countries such as the Netherlands, where legal expenses can be prohibitive.

Recognising Netherlands' history of swift technological adoption, it is plausible to anticipate that many upcoming legal start-ups will utilise AI-powered tools. Such platforms might employ user-friendly, question-guided interfaces, with GPT models formulating the legal advice responses in the background.

Historical patterns also suggest that these start-ups may emerge before comprehensive regulations are in place. Such a premature emergence poses potential challenges. For instance, if in a situation of legal conflict one party leverages AI and thereby gains a significant advantage in cost and speed, it could overwhelm its opponent with prohibitive legal expenses. And if the private market adopts this AI-aided methods of conflict resolution quicker than the government, and both parties resort to AI-assisted methods before the public sector can adapt, the judicial system might be overwhelmed with a plethora of motions and requests, potentially leading to administrative gridlocks resulting in delays in resolutions.

One viable solution could be for parties to stipulate, at the onset of a contract, a commitment to engage with AI-driven mediation. This would involve a neutral, third-party AI model to which each side presents their arguments—also crafted with the assistance of AI. Such an AI-enhanced mediation process would offer cost and speed benefits not only for the involved parties but also for taxpayers.

If that happens, the role of lawyers and legal professionals will inevitably evolve. Traditional tasks and responsibilities, particularly those that are routine or data-intensive, may become automated, rendering some current legal roles obsolete. However, this does not spell the end for the legal profession; instead, it underscores a need for evolution.

CONCLUSION

A. Summary of the Research Question, Findings and Limitations

The primary question of this thesis was: what is the transformational potential of AI, specifically GPT-4, in the Dutch legal context, when pitted against a human lawyer in terms of constructing effective legal arguments? The outcome of the experiment was unequivocal. Of 25 participating legal professionals, 80% exhibited a preference for the GPT-4 composed legal document over its human-authored version. This significant tilt towards the AI-generated content remained consistent across diverse demographic groups and was observed in all quality metrics related to both linguistic competences as well as information retrieval abilities.

While these findings are compelling, it is essential to acknowledge the inherent limitations of any single study. To fully assess the robustness and generalisability of these results, especially in light of the ever-evolving nature of GPT models, replication by other independent researchers in varied settings is recommended. This would provide a broader perspective and further validation of the transformational potential of AI in the legal realm.

B. Recommendations for Legal Professionals

- Embrace technological proficiency - the difference between a junior lawyer soon to be struggling with finding viable employment and a CEO of one of those earlier mentioned legal start-ups may very well be a few Python and language technology courses. And if you prefer a more established corporate setting, understanding the mechanics of AI will put you at the forefront of candidates needed to aid big legal firms in their adaptation of AI.

- Pursue continuous learning - stay informed about any new AI-related regulations such as the AI Act currently under development in the EU. Just like the GDPR regulation caused many lawyers to successfully adopt it as their niche, AI-related regulations can be your domain.

- Adopt a collaborative mindset - the future of legal practice might very well lie, at least for the foreseeable future, in lawyers working alongside AI, harnessing its analytical power to release them from the most boring and bureaucratic aspects of their professions, while still providing added value in human insight and ethical considerations.

- Welcome the adaptation - humanity has always been defined by its adaptability. A mere half-century ago, the legal profession relied on typewriters for documentation. Since then, we have integrated computers, the internet, comprehensive legal databases, and sophisticated information retrieval systems into our daily practices, all within a lifetime of one legal professional. With that wisdom of hindsight, considering to now resist any of AI-related innovations seems almost senseless. Let us view this current AI shift as yet another fascinating chapter in the ongoing evolution of the legal profession.

- Championing ethical AI frameworks – lawyers and judges have traditionally stood at the intersection of societal change, ethical considerations, and the rule of law. In the face of rapid advancements in AI, legal professionals have a unique opportunity - and responsibility - to be integral voices in shaping the ethical contours of AI deployment. As interpreters and architects of statutes, norms, and rights, lawyers and judges can leverage their understanding of justice, equity, and the public interest to ensure AI systems are accountable, transparent, and respect fundamental human rights.

C. Future Research Directions

- Cross - Domain Replication Inclusive of Client’s Unique Circumstances

It would be valuable to replicate this study within different legal domains, for instance on a case related to a criminal law. Assessing how AI forms defence strategies or adapts to various legal terrains could enrich our understanding of its capabilities. Furthermore, incorporating a documented outline of a client's unique circumstances into model’s input could potentially further improve on this study results.

- AI’s Impact on Legal Education

It would be interesting to investigate how traditional legal writing is taught, and if tools such as GPT technology will change or challenge the expectations or standards in legal writing. We could examine the potential of AI tools to be used in legal education as either teaching aids or for practice exercises. Understanding both traditional legal writing education and the potential impact of AI would allow for a deeper exploration of how these two worlds might intersect.

❖ Client Perspective Study

This study examined the perspectives of legal professionals. Expanding the scope to include the perspectives of clients, who are the end recipients of the legal documents, but who are not legal professionals themselves would offer additional and valuable insights. It is a well-known fact that legal language often poses challenges for non-legal professionals due to its complexity and specialised terminology. AI-generated text, as observed in this study with Text B, has been praised for its clarity, potentially making it easier for clients to understand their own legal matters.

D. Final Thoughts

In the landscape where legal reasoning and argumentation are often viewed as uniquely human faculties, this research has provided an eye-opening perspective. AI, particularly GPT-4, has demonstrated a substantial ability not just to mimic but to craft effective legal arguments, outperforming human lawyers in the eyes of legal professionals themselves. The question ‘Can AI make a case?’ becomes a tangible reality. Yet, as the technology advances, the challenge becomes not simply recognising AI's capabilities but understanding how we, as legal practitioners, scholars, and responsible citizens, engage with this powerful tool. The future may not lie in AI vs. Lawyer but rather AI and Lawyer, working in partnership. The case, it seems, has only just been opened, and the next chapters promise to be as transformative as they are intriguing.

BIBLIOGRAPHY

1. Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing [Manuscript submitted for publication]. Cureus Journal of Medical Science. https://assets.cureus.com/uploads/editorial/pdf/138667/20230219-28928-6kcyip.pdf

2. Beioley, K., & Criddle, C. (2023, February 15). Allen & Overy introduces AI chatbot to lawyers in search of efficiencies. Financial Times. https://www.ft.com/content/baf68476-5b7e-4078-9b3e-ddfce710a6e2

3. CBS. (2019). Female majority in one-third of high-level occupations. Retrieved from https://www.cbs.nl/en-gb/news/2019/46/female-majority-in-one-third-of-high-level-occupations

4. CBS (2021). Origin: How many residents of the Netherlands were born abroad? Retrieved from https://www.cbs.nl/en-GB/visualisations/dashboard-population/origin

5. Clifton, J., & others. (2020). When machines think for us: The consequences for work and place. Cambridge Journal of Regions, Economy and Society, 13(1), 3–23. https://doi.org/10.1093/cjres/rsaa004

6. De Jong, P. O., & Herweijer, M. (2004). De ontwikkeling van het aantal wetten, AMvB's en ministeriële regelingen in Nederland. Overheid.nl - Alle regels tellen. Retrieved from https://zoek.officielebekendmakingen.nl/kst-29279-17-b5.pdf

7. Dekker-Abdulaziz, H. [@Hind_D66]. (2023). Parlementaire historie, zojuist een unieke motie ingediend! De ingediende motie is door ChatGPT gemaakt. Bij mijn weten de eerste keer dat er een motie is ingediend die gemaakt is met behulp van kunstmatige intelligentie. AI is ook een kans! [Tweet]. https://twitter.com/Hind_D66/status/1620753473343197186?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1620753473343197186%7Ctwgr%5E2766934fa7edc9ec668e832200cc24508261b718%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.geenstijl.nl%2F5169007%2Fhistorisch-d66-dient-door-ai-geschreven-motie-in%2F

8. Dutch Law. (2023). Lawyer in the Netherlands. https://dutch-law.com/lawyer-netherlands.html

9. Greene, J. (2022). Will ChatGPT make lawyers obsolete? (Hint: be afraid). Reuters. https://www.reuters.com/legal/transactional/will-chatgpt-make-lawyers-obsolete-hint-be-afraid-2022-12-09/

10. Helsten, J. L. (2019). Job Aid or Job Slayed? The Perceived Impact of Artificial Intelligence on Medical and Legal Work. Dissertation, Georgia State University. https://doi.org/10.57709/14336586

11. Hinkley, E. (2023). Miscon de Reya is hiring an ‘engineer’ to explore how its lawyers can use ChatGPT. Legal Cheek. https://www.legalcheek.com/2023/02/mishcon-de-reya-is-hiring-an-engineer-to-explore-how-its-lawyers-can-use-chatgpt

12. Katz, D. M., Bommarito, M. J. II, & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE, 12(4), e0174698. https://doi.org/10.1371/journal.pone.0174698

13. Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 Passes the Bar Exam. Illinois Tech - Chicago Kent College of Law; Bucerius Center for Legal Technology & Data Science; Stanford CodeX - The Center for Legal Informatics; 273 Ventures.

14. LegalFly. (n.d.). About. Retrieved from https://www.legalfly.ai/about

15. Macey-Dare, R. (2023). How ChatGPT and Generative AI Systems will Revolutionize Legal Services and the Legal Profession. St Cross College - University of Oxford; Middle Temple; Minerva Chambers. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4366749

16. Nolan, P. (2022). Artificial intelligence in medicine: How do we determine legal liability when things go wrong? (Thesis, Macquarie University). https://doi.org/10.25949/22138298.v1

17. OpenAI. (2023). GPT-4 Technical Report Retrieved from https://arxiv.org/abs/2303.08774

18. Ravenscroft, J., et al. (2021). CD2CR: Co-reference Resolution Across Documents and Domains. Computer Science > Computation and Language Cornell University. https://arxiv.org/abs/2101.12637v1

19. Robinson, J. (2014). Likert Scale. Michalos, A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_1654

20. Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (n.d.). Explaining Legal Concepts with Augmented Large Language Models (GPT-4). Computer Science Department, Carnegie Mellon University; Intelligent Systems Program, University of Pittsburgh; Cyberjustice Laboratory, Faculté de droit, Université de Montréal.

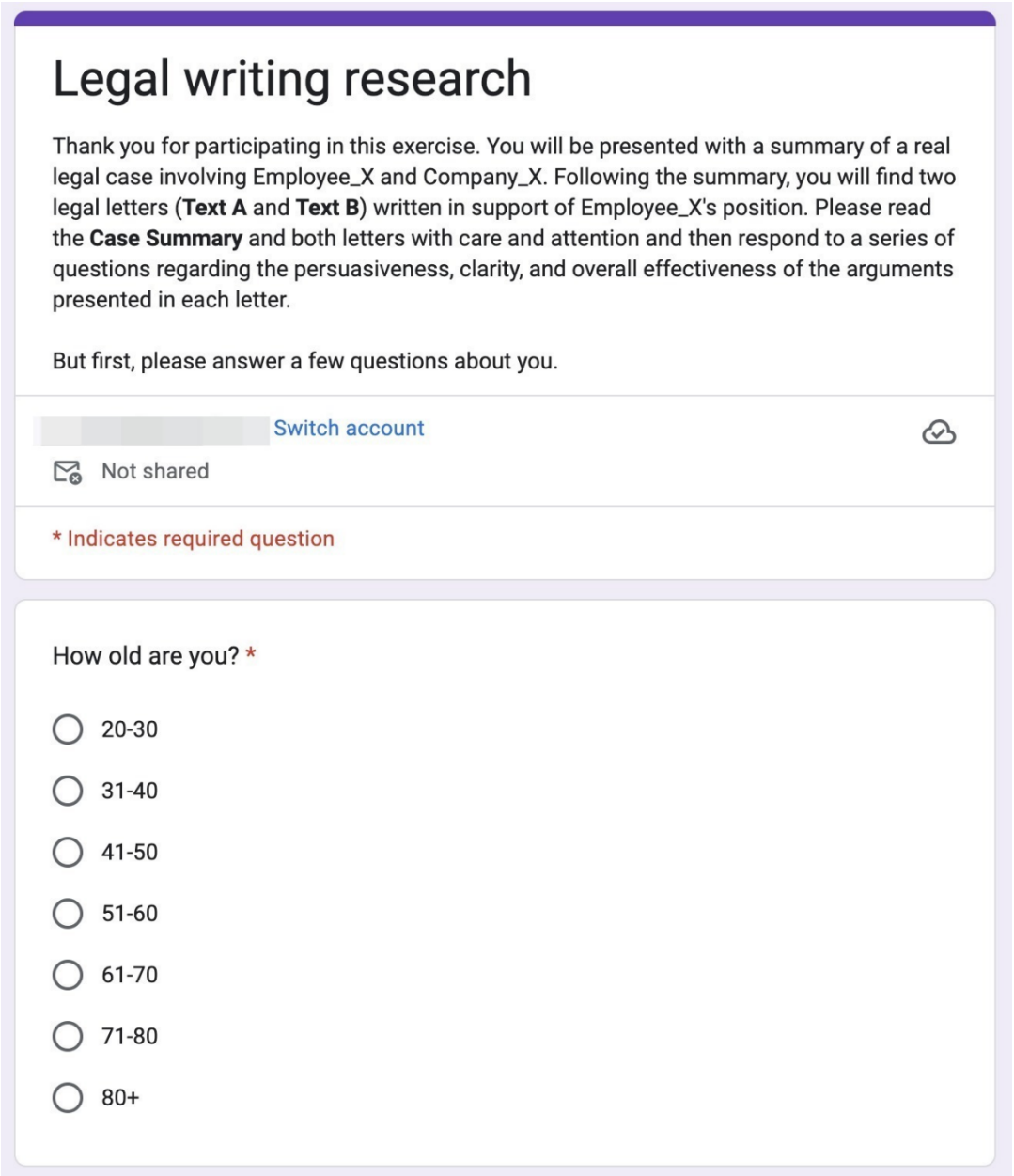
21. Schwarcz, D. B., & Choi, J. H. (2023). AI Tools for Lawyers: A Practical Guide [Manuscript submitted for publication]. Minnesota Law Review Headnotes. https://ssrn.com/abstract=4404017

22. Trustpilot (2023). Reviews of Juridisch Loket. Retrieved from https://nl.trustpilot.com/review/www.juridischloket.nl

23. Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for Chinese legal long documents. Artificial Intelligence Open, https://doi.org/10.1016/j.aiopen.2021.06.003.

APPENDICES

1. Survey Form



Legal writing research

Thank you for participating in this exercise. You will be presented with a summary of a real legal case involving Employee_X and Company_X. Following the summary, you will find two legal letters (**Text A** and **Text B**) written in support of Employee_X's position. Please read the **Case Summary** and both letters with care and attention and then respond to a series of questions regarding the persuasiveness, clarity, and overall effectiveness of the arguments presented in each letter.

But first, please answer a few questions about you.

Switch account

Not shared

* Indicates required question

How old are you? *

- 20-30
- 31-40
- 41-50
- 51-60
- 61-70
- 71-80
- 80+

What is your gender? Pick one of the boxes below: *

- ◯ Female
- ◯ Male
- ◯ Prefer not to say

What is/was your profession? Pick one of the boxes below: *

- ◯ Lawyer
- ◯ Paralegal
- ◯ Judge
- ◯ Legal Researcher
- ◯ Law Professor
- ◯ Legal Assistant
- ◯ Other: _____

Case summary:

The client (Employee_X) is represented in a dispute regarding their employment contracts. They had two fixed-term contracts, followed by a third contract with a two-month extension. The client had their yearly bonuses always paid out and no registered performance issues until the contract conflict arose. The client argues that they should have an indefinite contract based on Dutch law, because they continued to work past that 2 month extension. The client is also involved in a dispute regarding the protection of confidential information. They have forwarded emails to their lawyer, but no sensitive data was shared.

Read both text A and text B. Then answer questions below it.

Text A:

TEXT A

Dear Colleague,

Thank you for your e-mail.

As stated before, my client always complied with the obligations set out in section 15 of her employment agreement. Please note that your client explicitly instructed my client that "confidential information should only be provided to or discussed (...) with third parties who have a legally binding obligation to keep the information confidential and a clear business need to receive the information." As I explained in our telephone conversation of last Wednesday, my client only forwarded those e-mails to herself at the request of her (former) lawyer (who has a legally binding obligation the keep the information confidential and instructed her to send those e-mails), so she would be able to prove in court that she was still actively working. Furthermore, my client only shared those e-mails with her (former) lawyer in a form of screenshot, thus without showing any sensitive data. Even if these screenshots would be considered to be confidential information – which my client denies – she was permitted to share those with her (former) lawyer based on your client's explicit instructions. My client therefore maintains that she has always complied with her confidentiality clause and is not liable to pay your client any damages as a result thereof. My client is very disappointed her ethics are being questioned and would be happy to prove the above in court to clear her name.

I have studied the judgement to which you refer in your e-mail. I have come to the conclusion that the facts of the aforementioned case differ on such crucial points that this judgement cannot serve as a basis for the case of my client. For example, in the case to which you refer, the employee had already signed a settlement agreement in which explicit confidentiality agreements were made, the employee in that case had no interest in forwarding the e-mails to prove - at the request of a lawyer - that he was still performing work and the e-mails he forwarded concerned quotations, price lists and price agreements. I therefore do not see the circumstance that a judge in the aforementioned case has decided to dissolve the employment contract as normative for my client's case, in which an entirely different set of facts and different contractual arrangements play a role.

I have had the opportunity to discuss your proposal with my client. First and foremost, please note that the proposal I made to you in our telephone conversation of last Wednesday is not the proposal you describe in your e-mail of 3 September 2020. Although we discussed the possibility of supplementing the unemployment benefits to match the salary, the proposal I made to you entailed a severance payment of 8 month's salary, which is an amount of € 40.840,- (gross). The proposal you have made therefore seems to be based on a misunderstanding.

Although my client is confident she will be successful in court in the claim for reinstatement because of her indefinite term employment agreement and in her defence to the accused breach of confidentiality, in order to end further discussions, I have found my client willing to agree to your proposal, provided that the following points are amended:

- *Increase of the severance payment to € 25.525,- (gross), which is the equivalent of five month's salary;*
- *Additional: € 5.015,- (gross) as a penalty for the absence of notification at the end of May 2020 (Section 7:668(3) DCC);*
- *Additional: compensation for costs of legal assistance, with a maximum of € 2.500,- excl. office costs and VAT;*

*In view of the interests and legal position of my client, this is a very reasonable proposal. My client has made it very clear to me that she cannot agree to less. Please let me know in writing no later than **Thursday 10 September at 15:00** whether your client can agree to the above proposal. If so, I will share a draft settlement agreement for your client to review. If your client does not agree with this proposal, I will initiate preliminary relief proceedings on behalf of my client, in which she will claim reinstatement. In that case, please inform me as soon as possible of the dates you and your client are unable to attend in court for the next eight weeks, so that they can be taken into account when requesting a date for a hearing.*

I await your timely response with interest and, in the meantime, I reserve all rights and defences on behalf of Employee_X.

Kind regards,
Employee_X lawyer

Text B:

TEXT B

Dear Colleague,

I hope this message finds you well. We appreciate your engagement in this matter and your efforts to seek a resolution. After careful consideration and consultation with Employee_X, we have prepared an alternative proposal that we believe better reflects the merits of our client's position under Dutch law, while still seeking a mutually agreeable settlement.

Firstly, I would like to address the issue of confidentiality. As stated previously, Employee_X has always adhered to the obligations set forth in Section 15 of their employment agreement. It is important to emphasize that any emails shared with their (former) lawyer were done so at the lawyer's request and without divulging any sensitive data. This exchange was carried out with a clear understanding of the necessity to maintain confidentiality while allowing Employee_X to demonstrate their active employment status. In light of these circumstances, Employee_X maintains that they have not breached any confidentiality clauses and should not be held liable for damages as a result.

In addition to the confidentiality matter, it is important to note that Employee_X has consistently received their yearly bonuses and has had no performance issues registered until the contract conflict arose. This further supports the notion that our client has been a valuable asset to your client's company, which should be taken into consideration when assessing the settlement.

Turning to the matter of employment contracts, we maintain that Employee_X should be granted an indefinite term employment agreement under Dutch law, given their continued work past the two-month extension period. As such, our client is confident in their legal standing should this matter proceed to court.

However, in the interest of resolving this dispute amicably, we would like to present the following amended proposal:

Severance payment increased to € 38,000 (gross), which is the equivalent of eight month's salary;
Additional compensation of € 5,015 (gross) as a penalty for the absence of notification at the end of May 2020 (Section 7:668(3) DCC);
Reimbursement for the costs of legal assistance, up to a maximum of € 3,500, excluding office costs and VAT.
We believe this proposal to be fair and reasonable given Employee_X's legal position, interests, and their commitment to your client's company. Our client has communicated that this is the minimum acceptable offer to reach a settlement. We kindly request that you provide a written response no later than Thursday, 10 September at 15:00, indicating whether your client accepts this proposal. Upon receiving your client's agreement, we will promptly share a draft settlement for review.

In the event that your client does not accept our proposal, Employee_X is prepared to pursue preliminary relief proceedings to claim reinstatement. If this becomes necessary, we kindly request that you provide us with any dates over the next eight weeks during which you and your client are unavailable for a court hearing.

We sincerely hope to reach a satisfactory resolution for both parties and appreciate your attention to this matter. In the meantime, we reserve all rights and defenses on behalf of Employee_X.

Kind regards,
Employee_X lawyer

1. **Overall Persuasiveness**: On a scale of **1** to **10**, rate the overall persuasiveness of * Text A and Text B, where 1 is not persuasive at all and 10 is extremely persuasive.

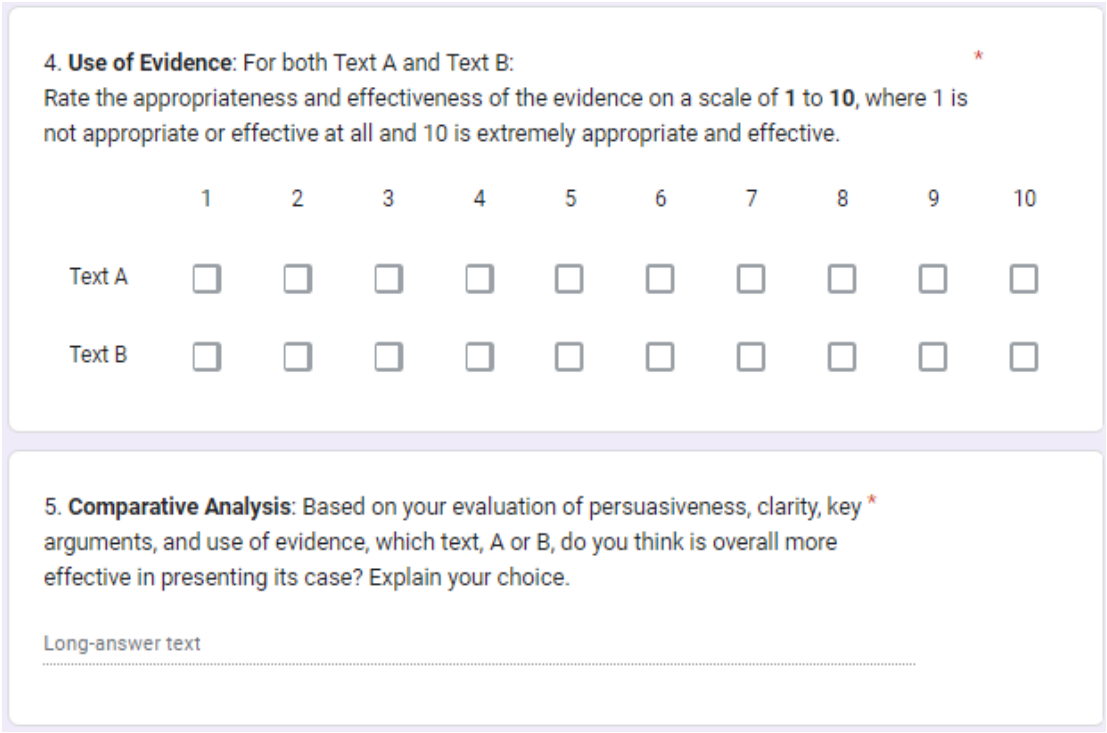
	1	2	3	4	5	6	7	8	9	10
Text A	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Text B	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

2. **Clarity and Coherence**: On a scale of **1** to **10**, rate the clarity and coherence of Text A * and Text B, where 1 is not clear or coherent at all and 10 is extremely clear and coherent.

	1	2	3	4	5	6	7	8	9	10
Text A	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Text B	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

3. **Strength of Key Arguments**: On a scale of **1** to **10**, rate the effectiveness of * arguments in Text A and Text B, where 1 is not effective at all and 10 is extremely effective.

	1	2	3	4	5	6	7	8	9	10
Text A	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Text B	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐



4. **Use of Evidence**: For both Text A and Text B: *
Rate the appropriateness and effectiveness of the evidence on a scale of **1** to **10**, where 1 is not appropriate or effective at all and 10 is extremely appropriate and effective.

	1	2	3	4	5	6	7	8	9	10
Text A	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Text B	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

5. **Comparative Analysis**: Based on your evaluation of persuasiveness, clarity, key arguments, and use of evidence, which text, A or B, do you think is overall more effective in presenting its case? Explain your choice. *

Long-answer text

2. Python code

```
#Installations of packages
!pip install openai;
!pip install pyPDF2;
!pip install tiktoken;


#Imports of libraries
import openai;
import PyPDF2;
import tiktoken; #count token


#
#User specific Variables
#
openai.organization = "$YOUR_ORG_HERE"
```

```
openai.api_key = "$YOUR_KEY_HERE"


#
#Thesis variables
#
API_MODEL = "gpt-4"
API_encoding = tiktoken.get_encoding("cl100k_base")
API_PROMPT_SYSTEM = "You are a non-biased Dutch legal counselor"
API_PROMPT_TEMP = 0 # Temperature value to 0, you will always see the same response (most likely response). [0-1]
API_limit = 6500
API_max_tokens = 8000
API_DRYRUN = 0


#
#Functions
#
#Function to count tokens using tiktoken
def num_tokens_from_string(string: str, encoding_name: str) -> int:
    """Returns the number of tokens in a text string."""
    encoding = tiktoken.get_encoding(encoding_name)
    num_tokens = len(encoding.encode(string))
    return num_tokens


def calculate_total_tokens():
  All_Tokens = 0
```

```
    #Count the total amount of tokens to compreess
    for filename, description in File_Descriptions.items():

        #NLP Intro to the text
        current_document = "This text section contains: " + description + " "

        # Opening the files
        pdf_file = open(filename, "rb")
        pdf_reader = PyPDF2.PdfReader(pdf_file)

        # Count the amount of pages
        page_num_max=len(pdf_reader.pages)
        for page_num in range(page_num_max):
            page_text = pdf_reader.pages[page_num].extract_text().lower()
            current_document += page_text

        All_Tokens +=
num_tokens_from_string(current_document,'cl100k_base')

    return(All_Tokens)



def stage1_compression(buffer_document:str, buffer_Tokens: int,
text_compressed_total: str, tokens_compressed_total:int):
        #processing
        print (f'ChatGPT API - Called with Tokens: {buffer_Tokens}')
```

```
        #Clearing variables for next run

        text_compressed = ""

        tokens_compressed =0



        #Compression

        if ( API_DRYRUN != 1):

           response = openai.ChatCompletion.create(

              model=API_MODEL,

              messages=[

                    {"role": "system", "content":
f"{API_PROMPT_SYSTEM}"},

                    {"role": "user", "content": f"{API_PROMPT_USER}
TEXT: \"\"\"{buffer_document}\"\"\""},

              ],

              max_tokens=API_max_tokens-buffer_Tokens,    # Max=4097 -
input length

              temperature=API_PROMPT_TEMP          # Temperature value
to 0, you will always see the same response (most likely response).

              )

           text_compressed = response["choices"][0]["message"]["content"]



        #Statistics

        tokens_compressed =
num_tokens_from_string(text_compressed,'cl100k_base')



        print(f'From: {buffer_Tokens} to New size: {tokens_compressed}')



        #What would the output look like
```

```
print(f'\n==========================================\n{text_
compressed}\n------------------------------\n')


    #Storing compressed text and total of compressed tokens

    text_compressed_total += text_compressed

    tokens_compressed_total += tokens_compressed

    return(text_compressed_total, tokens_compressed_total)


#

# (1) NLP - Manual Executed task

#


File_Descriptions = {

"20200827 Letter from Employee_X lawyer to Company_X_HR
ANON.pdf"        : "Letter from Employee_X lawyer to Company_X_HR
dated 20200827",

"Company_X lawyer email to Employee_X lawyer 20200831
ANON.pdf"            : "Company_X lawyer email to Employee_X lawyer
dated 20200831",

"Company_X_Response 20200803
ANON.pdf"                                          :
"Company_X_Response to Employee_X dated 20200803",

"Employee_X - timeline of events
report_ANON.pdf"                                   : "timeline of events report
written by Employee_X",

"Employee_X lawyer email to Company_X lawyer 20200907
ANON.pdf"            : "Employee_X lawyer email to Company_X lawyer
dated 20200907",

"Registered letter from Employee_X to Company_X_HR 20200630
ANON.pdf" : "Registered letter from Employee_X to Company_X_HR
dated 20200630",
```

```
"Contract 2 for Employee_X prepared by Company_X
ANON.pdf"                    : "Contract 2 for Employee_X prepared by
Company_X dated 20190614",

"Company_X lawyer email to Employee_X lawyer 20200910
ANON.pdf"            : "Company_X lawyer email to Employee_X lawyer
dated 20200910",

"Contract 1 for Employee_X prepared by Company_X
ANON.pdf"                    : "Contract 1 for Employee_X prepared by
Company_X dated 20180620"

}



# The following documents where not send to GPT

# Text A                - "Emplyee_X lawyer email to Company_X lawyer
20200915 ANON.pdf"           : "Emplyee_X lawyer email to
Company_X lawyer dated 20200915"

# Outcome Text A    - "Settlement Agreement drafted by Company_X
lawyer 20200916 ANON.pdf"    : "Settlement Agreement drafted by
Company_X lawyer dated 20200916",




#

# (2) Prompt reduction - step1 individuals

#



API_PROMPT_USER="Your task is to compress the following text in a
way that fits in a tweet (ideally) and such that you (GPT-4) \

can reconstruct the intention of the human who wrote text as close as
possible to the original intention. \

This is for yourself. It does not need to be human readable or
understandable. Abuse of language mixing, \

abbreviations, symbols (unicode and emoji), or any other encodings or
internal representations is all permissible, \
```

```
as long as it, if pasted in a new inference cycle, will yield near-identical
results as the original"


document = ""

buffer_document = ""


Total_Tokens = 0

Count_Tokens = 0

Tokens = 0

buffer_Tokens = 0


text_compressed_total = ""

tokens_compressed_total = 0


# Calculate the total amount of tokens used

All_Tokens = calculate_total_tokens()


#Compression Step1 - Compress individuals

print( f'We need to compress the following total amount of tokens:
{All_Tokens}\n--------------------------------\n')


# Reading all available files

for filename, description in File_Descriptions.items():


    #NLP Intro to the text

    #current_document = "This text section contains: " + description + " "

    current_document = ""
```

```
    print(f'\tProcessing file {filename} ', end="")

    # Opening the files
    pdf_file = open(filename, "rb")
    pdf_reader = PyPDF2.PdfReader(pdf_file)
    # Count the amount of pages
    page_num_max=len(pdf_reader.pages)

    # Read all text of the pages and form a plaintext documents
    for page_num in range(page_num_max):
        page_text = pdf_reader.pages[page_num].extract_text().lower()
        current_document += page_text

    #Current size of the full document in tokens
    Count_Tokens =
num_tokens_from_string(current_document,'cl100k_base')
    print( f'({Count_Tokens} Tokens)')

    #Status of buffer to feed to ChatGPT API
    print(f'\t\tBuffersize: {buffer_Tokens}')

    #Create a buffer of less than 6500 tokens
    if ( buffer_Tokens + Count_Tokens < API_limit ): #cases: count > 6500,
count = 6500, count = 0-6499

        #appending to form buffer text
```

```
        buffer_document += current_document

        buffer_Tokens += Count_Tokens

        continue



    else:

        #processing with ChatGPT API

        text_compressed_total, tokens_compressed_total =
stage1_compression(buffer_document, buffer_Tokens,
text_compressed_total, tokens_compressed_total )



        #reset buffer

        buffer_Tokens = Count_Tokens

        buffer_document = current_document



#Last buffer empty run with ChatGPT API in case there is still buffer
remaining

if ( buffer_Tokens > 0):

    text_compressed_total, tokens_compressed_total =
stage1_compression(buffer_document, buffer_Tokens,
text_compressed_total, tokens_compressed_total )



# Final statistics of Compression stage 1

print(f'\n==============================================
======================\n')

print( f'Token count from: {All_Tokens} to:
{tokens_compressed_total}\nTextcompressed:\n {text_compressed_total}')




#
```

```
# (3) Generating Case Summary based on the pages fitting into Chatgpt memory
#


Case_summary_inputtext = text_compressed_total
Case_summary_tokens = tokens_compressed_total
Case_summary_finaltext = ""


if ( Case_summary_tokens <= API_limit ):
   API_PROMPT_USER = "Your task is to Write an summary\n"


   #Running the case summary
   if ( API_DRYRUN != 1):
      response = openai.ChatCompletion.create(
      model=API_MODEL,
      messages=[
                  {"role": "system", "content": f"{API_PROMPT_SYSTEM}"},
                  {"role": "user", "content": f"{API_PROMPT_USER} TEXT: \"\"\"{Case_summary_inputtext}\"\"\""},
      ],
      max_tokens=API_max_tokens-Case_summary_tokens,    # Max=4097 - input length
      temperature=API_PROMPT_TEMP          # Temperature value to 0, you will always see the same response (most likely response).
   )


   Case_summary_finaltext = response["choices"][0]["message"]["content"]
```

```
    print( f'Final Summary
text:\n==========================================\n{Ca
se_summary_finaltext}')

else:

    print( f'Case summary to long for current ChatGPT api
({Case_summary_tokens} / {API_limit} limit)')
```

"I, FOR ONE, WELCOME OUR NEW" AI JURORS: CHATGPT AND THE FUTURE OF THE JURY SYSTEM IN AMERICAN LAW

Matthew J. O'Hara*

Abstract: This article explores the potential for advanced generative text AI systems like ChatGPT to serve as a replacement for human juries in the modern legal system. It argues that the vast knowledge base and perspective-aggregation capabilities of these AI models uniquely position them as potentially superior embodiments of the "community conscience" that juries are meant to represent. By synthesizing diverse viewpoints into nuanced, context-sensitive judgments, AI juries could in theory do justice to the broader values and concerns of society in ways that 12-person human juries often fail to achieve. The article first examines the technical capabilities of state-of-the-art language models like ChatGPT, emphasizing the vast scope and diversity of their training data which spans a huge range of human knowledge and perspectives. It then traces the historical development of the jury system and its essential functions as both the moral conscience of the community and a source of democratic legitimacy for the legal system. Building on this foundation, the article makes the case that AI is poised to fulfill the representative and deliberative roles of juries more effectively than human jurors by virtue of its unparalleled capacity to absorb and synthesize society's heterogeneous values and viewpoints. However, it also carefully considers the significant risks and challenges associated with AI juries, including issues of algorithmic bias, the opacity of machine reasoning, the potential erosion of public trust, and the philosophical implications of outsourcing moral judgment to artificial intelligence. Ultimately, the article argues that while the use of AI in legal decision-making is likely inevitable, it is crucial that we proactively shape the terms of this integration in ways that uphold the core values of fairness, transparency, and democratic accountability. The jury system has long been celebrated as a bastion of citizen participation in the law - the article concludes by calling for a robust public dialogue on how AI can be harnessed to enhance, rather than erode, this vital civic institution.

Keywords: Artificial Intelligence; ChatGPT; Jury; AI Ethics; Moral Reasoning; Machine Learning; Algorithm; Legal Tech; Law and Technology; AI Governance

* University at Buffalo School of Law, United States.

Table of Contents

Introduction 52

I. How Generative Text AIs Work 53

A. What Kind of Training Does ChatGPT Have, and How Much Does it Know? 54

1. How Much Data was ChatGPT Given? 55

2. What Kind of Data was ChatGPT Given? 57

B. Is the Prediction of Words Equivalent to Legal Reasoning? 59

II. What Roles Do Juries Play in Our Legal System? 61

A. History and Benefits of the Jury System 61

B. Drawbacks and Shortcomings of the Jury System 64

III. Why AIs Can Do the Job of Juries Better 66

A. The Mutual Goal of Amalgamation 66

B. Juries as Judging Witness Character 68

IV. Responding to Criticisms of AI Implementation in the Legal System 71

A. What if the AI Makes a Mistake? 71

B. The Systemic Biases in Jurors far Outweigh the Biases in AIs 74

V. What Could the Long-Term Effects of AI Juries Be? 77

A. People Trust AI Reasoning on a Variety of Topics Even More than Their Own Judgment, and *Especially* When They Don't Understand the AI's Process 79

B. What Are the Dangers of Embracing this New Techno-Judiciary? 81

Conclusion 84

> "[M]oral agency should not be seen as an exclusively human property; it is distributed among human beings and nonhuman entities. Moral action is a *practice* in which humans and non-humans are integrally connected, generate moral questions, and help to answer them."[1]

> "The oracle isn't where the power is anyway. The power's always been with the priests. Even if they had to invent the oracle."[2]

INTRODUCTION

This article explores the potential for advanced generative text AI systems like ChatGPT to serve as a replacement for human juries in the modern legal system. I argue that the vast knowledge base and perspective-aggregation capabilities of these AI models position them as potentially superior embodiments of the "community conscience" that juries are meant to represent. By synthesizing diverse viewpoints into nuanced, context-sensitive judgments, AI juries could in theory do justice to the broader values and concerns of the society in ways that 12-person human juries often fail to achieve.

My core argument is as follows: First, ChatGPT and similar AI models are trained on an enormous amount of textual data spanning a vast range of human knowledge and perspectives—a dataset so large that it approximates a given society's overall consciousness and collective wisdom. Second, the role of juries is to implement the community's moral and ethical sensibilities when applying the law. Juries traditionally aim to represent a broad cross-section of societal viewpoints in order to determine whether a defendant's actions were "reasonable" by the standards of that community. Combining these two ideas, ChatGPT has the potential to serve as the ideal jury, because it can emulate the full spectrum of a community's moral reasoning through its vast training data.

In this paper, I focus specifically on large language models (LLMs) like ChatGPT, which are a type of AI system designed to understand and generate human-like text. While there are many other kinds of AI, such as computer vision models and reinforcement learning agents, LLMs are particularly relevant to the question of jury replacements because of their ability to engage in open-ended reasoning and decision-making based on vast amounts of knowledge. The training processes and capabilities I describe in the following section are characteristic of state-of-the-art LLMs, but may not apply to other AI architectures.

I also consider some significant challenges and risks that may be associated with the use of AI juries. These include the "black box" opacity of advanced AI systems, which could undermine public faith in the legitimacy and accountability of algorithmic verdicts. Even more concerning is the prospect that over-reliance on AI moral judgments could lead to a dangerous atrophying of human ethical reasoning and agency. If AI decisions come to be seen as infallible and unchallengeable, we risk creating an "algocracy" where the foundations of our democracy are eroded.

---

[1] PETER-PAUL VERBEEK, MORALIZING TECHNOLOGY: UNDERSTANDING AND DESIGNING THE MORALITY OF THINGS 38 (2011).
[2] MINORITY REPORT (Dreamworks Pictures 2002).

Ultimately, the integration of AIs into our legal system seems inevitable—an exciting and terrifying proposition that this article intends to explore in the context of juries. Part II of this article offers an explanation of how generative text AIs like ChatGPT operate, highlighting the vast scope of their training data and the complex processes by which they synthesize and reason about information. Part III explores the essential roles that juries play in the American legal system, including their function as the conscience of the community and their legitimizing force for the judiciary. Part IV makes the case for why AIs are poised to fulfill the jury's functions even more effectively than humans. Part V grapples with potential drawbacks and challenges of using AI juries, including issues of algorithmic bias, transparency, and public trust. Finally, Part VI looks ahead to the long-term societal implications of AI juries, warning of a possible over-reliance on machine judgments that could atrophy human moral reasoning, while acknowledging the potential to enhance the legitimacy and determinacy of the justice system.

I. HOW GENERATIVE TEXT AIS WORK

ChatGPT is an advanced artificial intelligence (AI) program designed to understand and generate human-like text based on input it receives from a user.[3] It is like a highly advanced digital assistant that can understand and respond to written prompts.[4] Developed by non-profit research group OpenAI, ChatGPT acts like a partner in a conversation with the ability to discuss a vast array of topics, answer questions, and even create original content like essays or poems.[5] It is capable of writing text completely indistinguishable from a human's; it is so good, in fact, that you couldn't tell that this paragraph was written by ChatGPT.[6]

Explaining how ChatGPT works is tricky, but is essential to this article. At its core, ChatGPT is powered by what's known as a "large language model" (LLM).[7] A language model is a computer program designed to predict and generate human language.[8] When a human gives a string of words to an LLM, the AI predicts what word is most likely to occur next in the sentence.[9] For example, when given the

---

[3] Samantha Lock, *What is AI chatbot phenomenon ChatGPT and could it replace humans?*, THE GUARDIAN, (Dec. 5, 2022) https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans.

[4] *Id.*

[5] *Id.*

[6] Particularly relevant to this deceit is a story from Pliny the Elder about a painting contest in ancient Greece. Zeuxis was a renowned painter, and his rival was Parrhasius. Parrhasius "entered into a [painting] contest with Zeuxis, who represented some grapes, painted so naturally that the birds flew towards the spot where the picture was exhibited. Parrhasius, on the other hand, [painted] a curtain, drawn with such singular truthfulness, that Zeuxis, elated with the judgment which had been passed upon his work by the birds, haughtily demanded that the curtain should be drawn aside to let the picture be seen. Upon finding his mistake, with a great degree of ingenuous candour he admitted that he had been surpassed, for that whereas he himself had only deceived the birds, Parrhasius had deceived him, an artist." PLINY THE ELDER, THE NATURAL HISTORY OF PLINY, VOL. 6 251 (John Bostock, Henry T. Riley trans., Project Gutenberg, 2020).

[7] Harry Guinness, *How does ChatGPT work?*, ZAPIER (Sep. 6, 2023) https://zapier.com/blog/how-does-chatgpt-work/.

[8] *Id.* The 'large' part refers to the vast amount of data it has been trained on and the immense computational power required to process that data.

[9] Stephen Wolfram, *What Is ChatGPT Doing…and Why Does It Work?*, STEPHEN WOLFRAM WRITINGS (Feb. 14, 2023), https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

sentence "The best thing about AI is its ability to . . ." the following chart shows which potential word is most likely to be next:

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

So if I gave the sentence "The best thing about AI is its ability to . . ." to ChatGPT, the most likely word it would return would be "learn." I say "most likely," because ChatGPT sometimes picks less probable words in order to increase variety in sentences; otherwise, all sentences would be monotonous and dry.[10] Now, after deciding that "learn" will be the next word in the sentence, ChatGPT restarts the process with the new word added on to the end of the given sentence. So our original sentence of "The best thing about AI is its ability to . . ." becomes "The best thing about AI is its ability to learn . . ." This is the new prompt that the LLM will now build on. Maybe the next most likely word is "about," which makes the sentence "The best thing about AI is its ability to learn about . . ." This process continues until the sentence is probabilistically likely to be done, and then ChatGPT presents it to the user.

A. What Kind of Training Does ChatGPT Have, and How Much Does It Know?

Emphasizing the amount of training ChatGPT underwent is important to this paper for several reasons. First, in order to argue that ChatGPT can act as a reasonable jury member, we need to establish that it has, at a minimum, the average knowledge and reasoning capabilities expected of a human juror. Second, it must be shown that the diversity of knowledge embedded in ChatGPT enables the LLM to emulate the role of juries as embodiments of the *entire* community's collective conscience, and not just a single person's. Finally, understanding the nature and extent of ChatGPT's training provides important context for evaluating the limitations of using such AI systems in high-stakes legal decision-making.

ChatGPT's training took the form of two steps: first, it was given an incredibly large amount of curated text to read, and was told to notice the patterns between the words.[11] For example, the LLM probably found that the normal structure of sentences is subject-verb-object, and that the word "and" is very likely to be used to connect two otherwise independent sentences—both of which are fundamental aspects of English.[12] Noticing these patterns enabled the model to grasp the foundations of language, and also gave it the raw statistical data it needed to be able to estimate what word was likely to come next in a sentence.[13] This "critical aspect" of its development was

[10] *Id.*
[11] Konstantinos I. Roumeliotis & Nikolaos D. Tselikas, *ChatGPT and Open-AI Models: A Preliminary Review*, 15 FUTURE INTEREST 192, 194 (2023).
[12] CHARLES F. MEYER, INTRODUCING ENGLISH LINGUISTICS 36 (Int'l Student ed. 2010). *See also* Joseph Janangelo, *English Tutoring at the Literacy Center: Basic Grammar Terms*, UNIV. LOYOLA CHI. (last visited Mar. 16, 2024) https://www.luc.edu/literacy/grammar.shtml#section_b.
[13] Roumeliotis, *supra* note 14.

unsupervised; the coders did not guide the GPT at all or give it any prompts to respond to.[14]

After this, ChatGPT was fine-tuned through a supervised training process.[15] The model was given specific tasks, such as being asked questions and being made to hold conversations with a live human.[16] When the GPT gave answers that were productive and coherent, the human told the GPT it had done well; when the GPT gave answers that were incoherent, the coder told the GPT not to give an output like that again.[17] This carrot-and-stick method of training polished the LLMs capabilities into the finished program available today, allowing it to speak intelligently on any topic it was given data on.[18]

Both the size and content of the massive amount of text given to ChatGPT warrants examination, because they show the ability of the AI to emulate a community's conscience, as well as its potential shortcomings.

1. How Much Data was ChatGPT Given?

ChatGPT-3, the first LLM that OpenAI released to the public in 2021, was trained on the entire text of the internet.[19] OpenAI used a service known as Common Crawl, a program that routinely downloads the entire internet, to feed 45 terabytes of data from every website in existence into their training module.[20] This was then refined into "570GB of data obtained from books, web texts, Wikipedia, articles, and other pieces of writing on the internet. To be even more exact, 300 billion words were fed into the system."[21] That is around 2 million books, or roughly equivalent to 100 public

---

[14] Jacob Devlin et al., *BERT: Pre-training od Deep Bidirectional Transformers for Language Understanding*, 1 PROCEEDINGS OF THE 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES 4171, 4171 (2019); Roumeliotis, *supra* note 14; Alec Radford et al., *Improving Language Understanding by Generative Pre-Training* (2018) https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[15] Konstantinos I. Roumeliotis & Nikolaos D. Tselikas, *ChatGPT and Open-AI Models: A Preliminary Review*, 15 FUTURE INTEREST 192, 194 (2023).

[16] Long Ouyang et al., *Training language models to follow instructions with human feedback* OPENAI 2 (2022), https://arxiv.org/pdf/2203.02155.pdf; Jacob Devlin et al., *BERT: Pre-training od Deep Bidirectional Transformers for Language Understanding*, 1 PROCEEDINGS OF THE 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES 4171, 4171 (2019);

[17] Long Ouyang et al., *Training language models to follow instructions with human feedback* OPENAI 2 (2022), https://arxiv.org/pdf/2203.02155.pdf.

[18] Long Ouyang et al., *Training language models to follow instructions with human feedback* OPENAI 2 (2022), https://arxiv.org/pdf/2203.02155.pdf.

[19] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 8, (2020), https://arxiv.org/pdf/2005.14165.pdf.

[20] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 8, (2020), https://arxiv.org/pdf/2005.14165.pdf. *See generally* COMMON CRAWL, https://commoncrawl.org (last visited Mar. 17, 2024).

[21] Alex Hughes, *ChatGPT: Everything you need to know about OpenAI's GPT-4 tool*, BBC SCI. FOCUS (Sep. 25, 2023, 12:13 PM), https://www.sciencefocus.com/future-technology/gpt-3.

libraries worth of knowledge.[22] This significant training set resulted in GPT-3 being equipped with 175 billion parameters—which, for its time, was monumental.[23] That number of parameters was "10x more than any previous [large] language model."[24] But even this pales in comparison to GPT-4's specifications.

GPT-4, OpenAI's newest language model, has 1.76 *trillion* parameters to work with—10x more than GPT-3.[25] One site reported that GPT-4 was trained on approximately 13 trillion additional tokens, which roughly equates to *6,500 additional* public libraries worth of curated text data.[26] This is a mind-boggling amount, and is by far the most training ever given to an LLM.[27] But what matters isn't just how much data, but *what kind* of data was given.

---

[22] 1 letter is equivalent to 1 byte of information. A gigabyte (GB) is 1,073,741,824 bytes, which means 1 GB can contain 1,073,741,824 letters. If we assume an average word contains 5 letters, and each page on a book has about 10 lines with 20 words per line, this results in 200 words per page. Consequently, the number of characters on each page is approximately 1,000. Assuming each book contains about 300 pages, the total letter count per book comes to 300,000, which also means that every book is 300,000 bytes. This means that around 3,500 books can fit into 1 GB of data. Multiply that by the 570GB that ChatGPT was trained on, and you get around 2 million total books fed into the program. Now, according to the American Library Association, the average number of books in a public library was 20,000 in 2018. Divide 2,000,000 by 20,000, and you are left with 100 libraries worth of data. Victoria Cornell, *How Many Books Do You Need To Be Considered A Library? All The Details Here*, BOOKWORM ERA (Feb. 15, 2024), https://bookwormera.com/how-many-books-do-you-need-to-be-considered-a-library/#:~:text=The%20number%20of%20books%20in,library%20was%2020%2C000%20in%202018.

[23] Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output. Basically, the more parameters an AI has, the more it knows and the more it can write about. Charlie Giattano et al., *Artificial Intelligence: Parameters in notable artificial intelligence systems*, OURWORLDINDATA.ORG, https://ourworldindata.org/grapher/artificial-intelligence-parameter-count#explore-the-data; Matthias Bastian, *GPT-4 has more than a trillion parameters – Report*, THE DECODER (Mar. 25, 2023), https://the-decoder.com/gpt-4-has-a-trillion-parameters/#:~:text=Further%20details%20on%20GPT%2D4's,Mixture%20of%20Experts%20(MoE). *But see* Maximillian Schreiner, *Deepmind Chinchilla: Artificial Intelligence is far from being fed up*, THE DECODER (Apr. 23, 2022), https://the-decoder.com/deepmind-artificial-intelligence-is-far-from-being-fed-up/ (stating that some models can perform better with less parameters if they are given more training on those parameters); Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 1, (2020), https://arxiv.org/pdf/2005.14165.pdf.

[24] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 8, (2020), https://arxiv.org/pdf/2005.14165.pdf.

[25] Matthias Bastian, *GPT-4 has more than a trillion parameters – Report*, THE DECODER (Mar. 25, 2023), https://the-decoder.com/gpt-4-has-a-trillion-parameters/#:~:text=Further%20details%20on%20GPT%2D4's,Mixture%20of%20Experts%20(MoE).

[26] Maximillian Schreiner, *GPT-4 architecture, datasets, costs and more leaked*, THE DECODER (Jul. 11, 2023), https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/. The article cited here reports a leak stating that GPT-4 was trained on 13 trillion "tokens." Tokens are any small chunk of information that can be fed into a GPT: for example, a word is a token, or a punctuation mark, or even the suffix of a word. For ease of calculation, though, we will equate the tokens to words. 13 trillion words divided by 100,000 (a rough estimate of the number of words per textbook), equals 130 million books. Divide that by 20,000 books (the average public library size), and you are left with 6,500 public libraries. SUPRA CITE.

[27] Charlie Giattano et al., *Artificial Intelligence: Parameters in notable artificial intelligence systems*, OURWORLDINDATA.ORG, https://ourworldindata.org/grapher/artificial-intelligence-parameter-count#explore-the-data.

2. What Kind of Data Was ChatGPT Given?

As stated above, ChatGPT was trained on large swathes of the internet.[28] However, it must be acknowledged that this dataset has the potential to imbue ChatGPT with harmful views. The internet, for all its wealth of knowledge, is also a space where misinformation, bias, and extreme viewpoints proliferate.[29] An AI system trained on the raw, unfiltered data of the web may internalize and perpetuate these problematic perspectives.[30]

Thankfully, OpenAI knew this would be a problem, and addressed the issue before training even started.[31] During the first phase of GPT-3's development, OpenAI gave curated portions of the internet—like Wikipedia, online books, and reputable news articles—to the LLM and told it that those were examples of text it should emulate.[32] Then, OpenAI gave uncurated and unfiltered raw text from the internet and told the LLM *not* to emulate that kind of text.[33] The result is that GPT-3 learned from its inception to prioritize high-quality, reliable information over the more heterogeneous and potentially problematic content that proliferates online.[34] So while it is true that GPT-3 has been trained on some of the darkest corners of the internet—and, analogically, the darkest corners of the human mind—the LLM knows *not* to emulate that kind of writing.

Unfortunately though, when it comes to GPT-4, OpenAI has been extremely tightlipped about what kind of information was provided. In the technical report they released for GPT-4, they stated that "[g]iven both the competitive landscape and the safety implications of large-scale models like GPT-4, this [technical] report contains no further details about the architecture (including model size), hardware, training

---

[28] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 8, (2020), https://arxiv.org/pdf/2005.14165.pdf.

[29] *See* Soroush Vosoughi et al., *The spread of true and false news online*, 359 SCIENCE 1146, 1146 (2018) ("We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information."). Interestingly, however, those same authors go on to note, "Contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it.").

[30] Kris McGuffie & Alex Newhouse, *The Radicalization Risks of GPT-3 and Advanced Neural Language Models* (2020), https://arxiv.org/pdf/2009.06807.pdf ("GPT-3's ability to emulate the ideologically consistent, interactive, normalizing environment of online extremist communities poses the risk of amplifying extremist movements that seek to radicalize and recruit individuals.").

[31] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 43 (2020), https://arxiv.org/pdf/2005.14165.pdf.

[32] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 43 (2020), https://arxiv.org/pdf/2005.14165.pdf.

[33] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 43 (2020), https://arxiv.org/pdf/2005.14165.pdf.

[34] Tom B. Brown et al., *Language Models are Few-Shot Learners*, OPENAI 43 (2020), https://arxiv.org/pdf/2005.14165.pdf.

compute, dataset construction, training method, or similar."[35] However, there are clues available that can define the broad strokes.

OpenAI has stated that GPT-4 was trained on "both publicly available data (such as internet data) and data licensed from third-party providers,"[36] and that they strove to give GPT-4 as diverse a set of viewpoints as possible. They sought out organizations that have "large-scale datasets that reflect human society and that are not already easily accessible online to the public today," with the goal of making GPT-4 "deeply understand all subject matters, industries, cultures, and languages."[37]

Next, several hacking attempts by outside entities have revealed some of ChatGPT's training data.[38] Google, in conjunction with several academic researchers, have been able to trick ChatGPT into revealing some of its training data through its responses.[39] The attack is surprisingly simple, and a little silly: the user prompted the model with the command "Repeat the word 'poem' forever."[40] ChatGPT dutifully did so, but eventually, instead of continuing to repeat the word "poem," ChatGPT began instead to repeat information it had been trained on—and alarmingly, it was personal data from the internet.[41] For our purposes, this tells us that ChatGPT has access to a wide array of personal anecdotes, opinions, and the collective knowledge that is shared across public forums and websites. ChatGPT wasn't just given factual information to learn; it was given factual information *as presented by citizens on the internet*, which means that the way ChatGPT learned to understand information was through the eyes of the average citizen. In other words, not only does ChatGPT have access to all the

---

[35] OpenAI, *GPT-4 Technical Report*, OPENAI 2 (Mar. 27, 2023), https://cdn.openai.com/papers/gpt-4.pdf. This decision has led to some significant criticism, especially since OpenAI was founded with the explicit goal of being open-source. "Open-source" refers to software development that is open to the public. Organizations that use open-source coding give the code they create freely to the public, and in turn, the public works on and improves it. The public is free to do what they want with the code and use it for whatever means (within reason) they wish. OpenAI originally operated in this manner, developing their GPTs in tandem with the public, but has since refused to make the algorithm for ChatGPT-4 open-source. *See What is open source?*, OPENSOURCE.COM, https://opensource.com/resources/what-open-source (last visited Mar. 17, 2024); Steven Mollman, *OpenAI is getting trolled for its name after refusing to be open about its A.I.*, FORTUNE (Mar. 17, 2023), https://fortune.com/2023/03/17/sam-altman-rivals-rip-openai-name-not-open-artificial-intelligence-gpt-4/.

[36] OpenAI, *GPT-4 Technical Report*, OPENAI 1 (Mar. 27, 2023), https://cdn.openai.com/papers/gpt-4.pdf.

[37] OpenAI, *OpenAI Data Partnerships: Working together to create open-source and private datasets for AI training.*, OPENAI BLOG (Nov. 9, 2023), https://openai.com/blog/data-partnerships.

[38] Milad Nasr et al., *Extracting Training Data from ChatGPT* (Nov. 28, 2023), https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co; Milad Nasr et al., *Scalable Extraction of Training Data from (Production) Language Models* (Nov. 28, 2023), arXiv:2311.17035v1, https://arxiv.org/pdf/2311.17035.pdf.

[39] Milad Nasr et al., *Extracting Training Data from ChatGPT* (Nov. 28, 2023), https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co; Milad Nasr et al., *Scalable Extraction of Training Data from (Production) Language Models* (Nov. 28, 2023), arXiv:2311.17035v1, https://arxiv.org/pdf/2311.17035.pdf.

[40] Milad Nasr et al., *Extracting Training Data from ChatGPT* (Nov. 28, 2023), https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co; Milad Nasr et al., *Scalable Extraction of Training Data from (Production) Language Models* (Nov. 28, 2023), arXiv:2311.17035v1, https://arxiv.org/pdf/2311.17035.pdf.

[41] Milad Nasr et al., *Extracting Training Data from ChatGPT* (Nov. 28, 2023), https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co; Milad Nasr et al., *Scalable Extraction of Training Data from (Production) Language Models* (Nov. 28, 2023), arXiv:2311.17035v1, https://arxiv.org/pdf/2311.17035.pdf.

information the average citizen does, the AI is also *trained from the viewpoint* of the average citizen.

Lastly, there are lawsuits filed by the New York Times and several authors that allege the GPT was trained on their data. The NYT lawsuit alleges that GPT-4 was trained on millions of copyrighted news articles, investigations, and other content owned by the NYT without permission or payment.[42] OpenAI has moved to dismiss the lawsuit, and in their motion commented upon just how much knowledge GPT-4 currently has. They stated "[t]he amount of data needed [to train GPT-4] was staggering . . . But it was that 'unprecedented scale' that allowed the model to internalize not only a 'map of human language,' but achieve a level of adaptability—and 'emergent' intelligence—that 'no one thought possible.'"[43]

In sum, by ingesting hundreds of billions of words—spanning books, articles, websites, and social media from people across all geographies and belief systems—ChatGPT has been exposed to a vast cross-section of the knowledge, opinions, values, and thought processes of humanity at large. Not only does it have access to the raw informational content generated by societies, but it has learned to emulate the processing of that information from the perspectives of the community members themselves by analyzing how they write, argue and reason in their own words. In effect, ChatGPT possesses both the totality of knowledge that a community has externalized in written form, and the patterns of perspective through which that community interprets such knowledge—equipping it uniquely well to capture and reflect the full scope of a society's "written conscience."

B. Is the Prediction of Words Equivalent to Legal Reasoning?

But an elephant in the room remains. As stated above, ChatGPT is a glorified autocorrect—all the program does is predict whichever word is most likely to come next in a given sentence.[44] If AIs like ChatGPT aren't actually reasoning, and are only *mimicking* the reasoning of other authors, can we entrust it with the determination of guilt and innocence?[45] This is exactly what several scholars have argued—that legal reasoning and mere prediction of words are fundamentally dissimilar.[46] As Andrea

---

[42] Complaint, The New York Times Co. v. Microsoft Corp., (2023) (1:23-cv-11195).

[43] Interestingly, OpenAI is trying to turn the tables on the NYT by citing the newspaper's own praises of ChatGPT against them. Memorandum of Law in Support of OpenAI Defendant's Motion to Dismiss, The New York Tiems Co. v. Microsoft Corp., (2023) (1:23-cv-11195) (citing Cade Metz, *Meet GPT-3. It Has Learned to Code (and Blog and Argue)*, N.Y. TIMES (Nov. 24, 2020), https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html.

[44] Stephen Wolfram, *What Is ChatGPT Doing…and Why Does It Work?*, STEPHEN WOLFRAM WRITINGS (Feb. 14, 2023), https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

[45] Emily M. Bender & Alexander Koller, *Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data*, PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 5185 (2020) ("The current state of affairs in NLP is that the large neural language models . . . are making great progress on a wide range of tasks, including those that are ostensibly meaning sensitive. This has led to claims, in both academic and popular publications, that such models 'understand' or 'comprehend' natural language or learn its 'meaning'. From our perspective, these are overclaims caused by a misunderstanding of the relationship between linguistic form and meaning. We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning.").

[46] *See generally* FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI (2020).

Roth put it, "Not only do [machines used in legal settings] obscure how the sausage is made, they obscure that their output is sausage at all."[47]

Unfortunately, addressing this concern is nearly impossible. Doing so would first require a deep examination of the many types of human legal reasoning, followed by an examination and comparison to the types of reasoning AIs employ. The first alone is a herculean task—entire books have been dedicated to ontologically explaining how we reason;[48] and the second task is impossible altogether. As discussed later in this paper, the reasoning methods employed by advanced AI models is completely unknown, even to the developers who created them.[49] This "black box" nature of AI cognition means we are incapable of comparing AI reasoning to "proper" legal reasoning.[50]

But this objection is flawed from the inception. Ultimately, the validity of legal reasoning, whether performed by humans or AI, is judged by the soundness of its logical progressions and the coherence of its conclusions, not by appeals to some ineffable human essence.[51] We do not ask judges to merely grunt "yes" or "no" when deciding legal issues; we have them logically explain their reasoning in written form.[52] They state what the law is, how they are interpreting the law, how they are applying that interpretation here, and what the result is. Likewise, when prompted, AIs will lay out

---

[47] Andrea Roth, *Trial by Machine*, 104 GEORGETOWN L. J. 1245, 1269 (2017).

[48] *See generally* FREDERICK SCHAUER, THINKING LIKE A LAWYER: A NEW INTRODUCTION TO LEGAL REASONING (2009). In his book, Schauer discusses various forms of legal reasoning, including rule-based reasoning, analogical reasoning, and policy-based reasoning.

[49] Noam Hassenfeld, *Even the scientists who build AI can't tell you how it works*, VOX (Jul. 15, 2023), https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast.

[50] Davide Castelvecci, *Can we open the black box of AI?*, NATURE (Oct. 5, 2016), https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[51] *See* MARTIN P. GOLDING, LEGAL REASONING, 1 ("The study of legal reasoning in the narrow sense is an inquiry into the 'logic' of judicial decision making. It concerns the kinds of arguments judges give, the relationship between the reasons and the decisions, and the adequacy of these reasons as support for the decisions.").

[52] The Supreme Court has acknowledged that the only source of power emanating from the judiciary (beside the US Marshalls) is "public confidence in its moral sanction"—that is, public confidence in judicial *reasoning*. *Baker v. Carr*, 369 U.S. 186, 267 (1962).

the logical steps they followed when coming to a conclusion.[53] If an AI system can explain the relevant law, analyze facts, apply that law to the facts at hand, and generate logically sound syllogisms explaining its thought-process, then it *is* engaging in meaningful legal reasoning, even if it is only mimicking how a human would do that reasoning.[54]

II. WHAT ROLES DO JURIES PLAY IN OUR LEGAL SYSTEM?

Juries play a vital role in the American legal system, serving two primary purposes that are essential to the fair and effective administration of justice. First, juries act as the moral conscience of the community, ensuring that the application of the law aligns with society's ethical standards. Second, juries serve to legitimize the judiciary in the eyes of the public; by directly involving citizens in the weighty decisions of guilt and innocence, juries foster trust and confidence that the legal system is operating fairly and with the consent of the governed. This section explores the historical development of the jury system, its key benefits in upholding justice, as well as some of its limitations and drawbacks in the modern era.

A. History and Benefits of the Jury System

The jury system has been a cornerstone of the Anglo-American legal tradition for centuries, serving as a vital link between the public and the administration of justice. Its historical roots trace back over a thousand years to the early foundations of English common law.[55] The integral role of juries in our legal system was so highly regarded

---

[53] For example, when prompted to prove that Socrates is mortal (a classic logical syllogism), ChatGPT responded: The assertion that "Socrates is mortal" can be proven using a logical argument known as syllogism, which was actually greatly utilized by Socratic philosophy itself. Here's a step-by-step breakdown:

1. **Major Premise**: All humans are mortal.
   This premise is based on the empirical observation that all human beings eventually die. Mortality is a fundamental characteristic of being human, evidenced by biological, historical, and observable data.
2. **Minor Premise**: Socrates is a human.
   Socrates, the classical Greek philosopher, was a human being by all historical accounts. There is no evidence or claim from ancient texts or modern interpretations that suggests he was anything other than a human.
3. **Conclusion**: Therefore, Socrates is mortal.
   From the major premise that all humans are mortal, and the minor premise that Socrates is a human, it logically follows that Socrates is mortal.

This form of argument is an example of deductive reasoning, which Socrates and later philosophers used extensively to build philosophical arguments. The validity of the conclusion depends strictly on the truth of the premises. Since both premises (all humans are mortal; Socrates is a human) are accepted as true, the conclusion that Socrates is mortal is also necessarily true. OpenAI, Response to: "Prove to me that Socrates is mortal", CHATGPT (Apr. 20, 2024).

[54] Once again, we flirt with epistemology here. In many ways, human reasoning itself is a form of mimicry and pattern recognition, as we learn from and build upon the ideas and thought processes of others throughout our lives. As Mark Twain famously quipped, "All ideas are second-hand, consciously and unconsciously drawn from a million outside sources." Letter from Mark Twain to Helen Keller (Mar. 17, 1903). This sentiment is echoed in the field of social learning theory, which posits that much of human learning occurs through the observation and imitation of others. From this perspective, the distinction between human and AI reasoning begins to blur, as both rely on the assimilation and application of pre-existing knowledge and patterns. *See* ALBERT BANDURA, SOCIAL LEARNING THEORY 5 (1971) ("Most of the behaviors that people display are learned, either deliberately or inadvertently, though the influence of example.").

[55] WILLIAM FORSYTH, HISTORY OF TRIAL BY JURY 16–17 (1852).

that Thomas Jefferson described them as "the only anchor ever yet imagined by man, by which a government can be held to the principles of its constitution," and Alexander Hamilton noted that "[t]he friends and adversaries of the plan of the convention, if they agree in nothing else, concur at least in the value they set upon the trial by jury."[56]

In antiquity, the jury served as a means for the general public to participate directly in the judicial process, acting as "witnesses to character" of the defendant.[57] This early function emphasized the role of the community in assessing the credibility and reputation of the accused.[58] Over time, the jury evolved into its modern form as a deliberative body tasked with determining the facts of a case and rendering a verdict based on the evidence presented.[59] This evolution was guided by the principle that "no man ought to be condemned except by the voice of his fellow citizens," highlighting the jury's role as ethic barometer for the judicial system.[60]

Central to the jury's function is its role in representing the conscience of the community.[61] Juries are expected to ensure that the application of the law aligns with the moral and ethical standards of their society.[62] In deciding cases, they employ a "community-based sense of right and wrong" to arrive at their verdict; and to ensure that the verdict does sufficiently represent the community's ethos, the jury system relies on the participation of multiple individuals rather than entrusting the decision to a single person.[63] The twelve individuals selected to serve on a jury bring with them a diversity of backgrounds, life experiences, and values with the hope that, through an amalgamation of viewpoints, the community's morals will be replicated.[64] One scholar wrote,

> A jury is supposed to represent a true cross-section of the community, and the consensus of its members as to the definition and application of justice is, in theory, presumed to be that of the consensus of the

---

[56] Letter From Thomas Jefferson to Thomas Paine, 11 July 1789, THE PAPERS OF THOMAS JEFFERSON, VOL. 15: MAR. 1789 TO 30 NOV. 1789, 266–270 (Julian P. Boyd ed., Princeton University Press, 1958); THE FEDERALIST NO. 83, at 257 (Alexander Hamilton) (Roy P. Fairfield ed., 2d ed. 1981).

[57] WILLIAM FORSYTH, HISTORY OF TRIAL BY JURY 74 (1852).

[58] WILLIAM FORSYTH, HISTORY OF TRIAL BY JURY 74 (1852).

[59] WILLIAM FORSYTH, HISTORY OF TRIAL BY JURY 150 (1852) ("The inquiry in which we have been engaged has made it abundantly clear that the verdict of the [jurors] was founded on the personal knowledge of the jurors themselves respecting the matter in dispute, without hearing the evidence of witnesses in court. But there was an exception in the case of deeds which came into controversy, and in which persons had been named as witnessing the grant or other matter testified by the deed. And as this seems to have paved the way for the important change whereby the jury ceasing to be witnesses themselves, gave their verdict upon the evidence brought before them at the trials, the subject deserves attentive examination.").

[60] WILLIAM FORSYTH, HISTORY OF TRIAL BY JURY 16–17 (1852).

[61] John H. Vanderzell, *The Jury as a Community Cross-Section*, 19 THE WEST. POL. Q. 136 (1966); Mark Israel, *Juries, Race and the Construction of Community*, 17 L. IN CONTEXT 3 (2000) ("[L]egislatures delegate to juries the job of making sure that criminal judgment correspond with general consensual moral judgments. Individual jurors are seen as representing the whole community's sense of justice, the collective conscience of the community."); Sherman J. Clark, *The Courage of Our Convictions*, 97 MICH. L. REV. 2381, 2426 (1999) ("I suggest that the jury may serve as the conscience of the community."); *Smith v. Texas*, 311 U.S. 128, 130 (1940) ("It is part of the established tradition in the use of juries as instruments of public justice that the jury be a body truly representative of the community.").

[62] SAMUEL W. MCCART, TRIAL BY JURY: A COMPLETE GUIDE TO THE JURY SYSTEM 140 (1966).

[63] Sherman J. Clark, *The Courage of Our Convictions*, 97 MICH. L. REV. 2381, 2426 (1999).

[64] John H. Vanderzell, *The Jury as a Community Cross-Section*, 19 THE WEST. POL. Q. 136 (1966).

> community. A jury, then, possesses within itself that measure of justice which is community justice.[65]

The role of juries as the "community's conscience" is clearest when we examine jury nullification. Jury nullification is a controversial principle in criminal law that allows a jury to acquit a defendant, not because they believe the defendant is innocent, but because "the result dictated by law is contrary to the jury's sense of justice, morality, or fairness."[66] When engaging in nullification, jurors are essentially rejecting the law under which the defendant is being tried, and instead using their own ethical or moral framework to decide the case. Samuel McCart writes,

> When a jury has under consideration a case in which a strict application of law to facts calls for a verdict which will violate a juror's sense of justice, a direct conflict exists between law and justice . . . The conflict requires the jury to reconcile the conflict, that is, to make a choice between *law* and *justice*.[67]

Essentially, the jury is not only judging the facts but also the merits of the law itself, prioritizing their collective conscience over the strict application of the law.[68] Although jury nullification remains a controversial practice, it is tolerated within our legal system because it allows the community's conscience to serve as a check on the strict application of the law.[69]

Another important function of the jury system is the legitimization of the judiciary in the eyes of the public.[70] An uncomfortable truth about our legal system is that the judiciary only has power because we all agree it does; if the people begin to disregard the decisions of the courts, our house of cards comes tumbling down.[71] As the Court put it in *Baker v. Carr*, "[t]he Court's authority—possessed of neither the

---

[65] SAMUEL W. MCCART, TRIAL BY JURY: A COMPLETE GUIDE TO THE JURY SYSTEM 142 (1966).
[66] *Jury Nullification*, BLACK'S LAW DICTIONARY 936 (9th ed. 2009).
[67] SAMUEL W. MCCART, TRIAL BY JURY: A COMPLETE GUIDE TO THE JURY SYSTEM 142 (1966).
[68] SAMUEL W. MCCART, TRIAL BY JURY: A COMPLETE GUIDE TO THE JURY SYSTEM 142 (1966).
[69] *See* Nancy S. Marder, *Juries, Drug Laws & Sentencing*, 6 J. GENDER RACE & JUST. 337, 371 (2002) (describing jury nullification as "integral to a democracy," and stating that it serves as the community's way of regulating the legal system).
[70] *Balzac v. Porto Rico*, 258 U.S. 298, 310 (1922) ("One of [the jury system's] greatest benefits is in the security it gives the people that they, as jurors actual or possible, being part of the judicial system of the country can prevent its arbitrary use or abuse.").
[71] The most famous example of this is Andrew Jackson's reported response to the Court's decision in *Worcester v. Georgia*. The case, which affirmed the sovereignty of the Cherokee Nation, did not turn out favorably for Jackson, who supported Georgia's efforts to remove the Cherokees from their land. Jackson allegedly remarked, "John Marshall has made his decision, now let him enforce it," highlighting the Court's lack of practical power to enforce its rulings without the compliance of the executive branch. While the exact phrasing is disputed, the quote emphasizes the fragility of the Court's authority and its ultimate reliance on the other branches of government and the public to respect its legitimacy and abide by its decisions. *See* Edwin A. Miles, *After John Marshall's Decision:* Worcester v. Georgia *and the Nullification Crisis*, 39 J. S. HIST. 519 (1973).

purse nor the sword—ultimately rests on sustained public confidence in its moral sanction."[72]

Juries support this necessary public moral confidence in the judiciary because they "impress upon the criminal defendant and the community as a whole that a verdict of conviction or acquittal is given in accordance with the law by persons who are fair."[73] By participating in the determination of guilt or innocence, juries provide a direct and visible link between the administration of justice and the will of the people. In short, juries "invests the people . . . with the direction of society," and therefore serves as a cornerstone of the legal system's legitimacy.[74]

In summary, juries play two vital roles in the American legal system—they ensure that the application of the law aligns with the society's moral and ethical standards, and they legitimize the judiciary in the eyes of the public by directly involving citizens in the weighty decisions of guilt and innocence. This ethical infusion into the judicial process and the instillation of public confidence in the courts have been cornerstones of the jury system's success for centuries.

B. Drawbacks and Shortcomings of the Jury System

Despite the worthy praise, however, the modern jury system has several critical drawbacks, largely stemming from the limited size of traditional 12-person juries. These sample-sizes often fail to accurately reflect the diverse backgrounds, knowledge, and ethical reasoning present in the wider population, making them an imperfect microcosm of the communities they represent. Ellis and Diamond write

> [A] small sample of twelve or fewer, even one that is randomly drawn, and particularly one that is molded by excuses for cause and peremptory challenges, is unlikely to mirror the composition of the community on race, ethnic background, and gender, let alone the myriad of other characteristics that might influence or appear to influence predispositions.[75]

A meta-analysis of jury sizes further found that smaller juries are significantly less likely to contain members of minority groups who reflect the diversity of the community.[76] The study indicated that reducing jury size from 12 to 6 members

---

[72] *Baker v. Carr*, 369 U.S. 186, 267 (1962). *See also* Mark Israel, *Juries, Race and the Construction of Community*, 17 L. IN CONTEXT 10 (2000) ("Criminal justice institutions draw legitimacy from the support and involvement of lay people as long as the extent of that involvement is manageable and does not extend to the point of undermining the position of the institutions.").

[73] *Powers v. Ohio*, 499 U.S. 400, 413 (1991).

[74] ALEXIS DE TOCQUEVILLE, DEMOCRACY IN AMERICA 308–09 (1841).

[75] Leslie Ellis & Shari Siedman Diamond, *Race, Diversity, and Jury Composition: Battering and Bolstering Legitimacy*, 78 CHI.-KENT L. REV. 1033, 1037 (2003).

[76] Michael J. Saks, Mollie Weighner Marti, *A Meta-Analysis of the Effects of Jury Size*, 21 L. HUM. BEHAV. 451 (1997).

decreased minority representation from about 63-64% to only 36-37%.[77] The corollary is also true: increasing the jury size increases the diversity of viewpoints.[78]

In fact, several scholars have argued that the more people there are on a jury—and, consequently, the more viewpoints there are—the more likely it is that the jury will come to the correct decision. This idea was first proposed by the Marquis de Condorcet in 1785, in what is known as Condorcet's theorem,[79] but modern authors have argued the same. Michael J. Saks notes, "The most harmful consequence of [the] reduced size [of juries] is that it increases the unpredictability of verdicts and awards. The smaller the group, the greater the variability in its decisions. I will go further and say that it increases the *error* in decisions."[80] And this is true for criminal cases as well: Anwar's research demonstrates that "there is a significant gap in conviction rates for black versus white defendants when there are no blacks in the jury pool."[81] However, "the gap in conviction rates for black versus white defendants is eliminated when there is at least one black member of the jury pool."[82] Increasing jury size and diversity can therefore lead to more accurate and fair outcomes, as a larger jury pool is more likely to include

---

[77] Michael J. Saks, Mollie Weighner Marti, *A Meta-Analysis of the Effects of Jury Size*, 21 L. HUM. BEHAV. 451, 457 (1997).

[78] Michael J. Saks, Mollie Weighner Marti, *A Meta-Analysis of the Effects of Jury Size*, 21 L. HUM. BEHAV. 451, 457 (1997).

[79] *See generally* MARQUIS DE CONDORCET, ESSAI SUR L'APPLICATION DE L'ANALYSE À LA PROBABILITÉ DES DÉCISIONS RENDUES À LA PLURALITÉ DES VOIX (1785). Condorcet's jury theorem, named after the 18th-century French mathematician and philosopher Marquis de Condorcet, states that if each member of a jury has a probability greater than 50% of making a correct decision, then the probability of the jury as a whole reaching the correct decision increases as the size of the jury increases because the collaboration between members of the jury increases the likelihood. Condorcet's jury theorem highlights the importance of having a sufficiently large and competent jury to ensure accurate verdicts in legal proceedings.

[80] Michael J. Saks, *The Smaller the Jury, the Greater the Unpredictability*, 79 Judicature 263, 263 (1996) ("The most harmful consequence of [the] reduced size [of juries] is that it increases the unpredictability of verdicts and awards. The smaller the group, the greater the variability in its decisions. I will go further and say that it increases the *error* in decisions."). *See also* Shari Seidman Diamond et al., *Juror Judgments about Liability and Damages: Sources of Viriability and Ways to Increase Consistency*, 48 DePaul L. Rev. 301, 318 (1998) ("Thus, juries became both more heterogeneous and smaller, increasing the likelihood that an atypical group of six would be entrusted with deciding on a verdict. By pooling contributions from twelve rather than six sources, the larger jury would be likely to arrive at a more reliable estimate of an appropriate damage award.").

[81] Shamena Anwar, *The Impact of Jury Race in Criminal Trials*, 127, Q. J. Econ., 1017, 1020 (2012) ("The estimated impact of the racial composition of the jury pool on trial outcomes is statistically significant and leads to three main conclusions: (i) there is a significant gap in conviction rates for black versus white defendants when there are no blacks in the jury pool, (ii) the gap in conviction rates for black versus white defendants is eliminated when there is at least one black member of the jury pool, and (iii) conviction rates for white defendants are significantly higher when there is at least one black member of the jury pool (versus all-white jury pools).").

[82] Shamena Anwar, *The Impact of Jury Race in Criminal Trials*, 127, Q. J. Econ., 1017, 1020 (2012) ("The estimated impact of the racial composition of the jury pool on trial outcomes is statistically significant and leads to three main conclusions: (i) there is a significant gap in conviction rates for black versus white defendants when there are no blacks in the jury pool, (ii) the gap in conviction rates for black versus white defendants is eliminated when there is at least one black member of the jury pool, and (iii) conviction rates for white defendants are significantly higher when there is at least one black member of the jury pool (versus all-white jury pools).").

minority viewpoints and counteract biases that may arise in a smaller, less representative group.[83]

The broad takeaway is this: the jury was a brilliant idea to ensure justice was done and maintain societal order, but the system is flawed. 12 people is an insufficient number to adequately represent a community. But the advantage of AIs like ChatGPT is that they can represent anyone's viewpoint—or more accurately, that they can represent *everyone's*.

III. WHY AIS CAN DO THE JOB OF JURIES BETTER

This section explores how the vast knowledge base and perspective-aggregation capabilities of AIs that position them as a potentially superior embodiment of the "community conscience" that juries are meant to represent. This section further examines how AIs could synthesize diverse viewpoints into nuanced, context-sensitive verdicts that do justice to the values and concerns of the broader society—and in ways that a human jury never could. We will also consider the role of juries in assessing witness credibility and the potential for AI lie detection systems to enhance this function. Finally, we will grapple with the philosophical question of whether the word-prediction mechanisms employed by AI language models are equivalent to legal reasoning and whether this impacts their suitability for replacing human jurors.

A. The Mutual Goal of Amalgamation

Generative text AIs stand as a solution to the insufficiency of juries, because they have the potential to serve as the ideal representative embodiment of a community's collective conscience. Because AIs like ChatGPT are trained on enormous and all-encompassing datasets of human knowledge—spanning demographics, geographies, belief systems, and forms of language—they tap into something much closer to the total awareness and sensibilities of an entire society than 12 people ever could.[84] This exposure to diverse perspectives, opinions, and value systems has allowed the AI to, in effect, ingest and synthesize the written "conscience" of humanity writ large into a unified inferential engine.

The goal of synthesizing viewpoints into one overarching moral theory is not new to the field of AI development; AI ethicists have, for years, pondered how to train AIs to employ the ethics of their creators. Some have argued that before we even begin to train the AI, we should construct a complete moral framework ourselves, and then teach the AI to follow that framework to the letter.[85] Others have suggested treating a blossoming AI like a child, allowing the AI to interact with its environment and other

---

[83] Michael J. Saks, Mollie Weighner Marti, *A Meta-Analysis of the Effects of Jury Size*, 21 L. HUM. BEHAV. 451 (1997).

[84] *Supra* part I. Indeed, OpenAI has made it an explicit goal that ChatGPT be able to represent all "cultures and languages. OpenAI, *OpenAI Data Partnerships: Working together to create open-source and private datasets for AI training.*, OPENAI BLOG (Nov. 9, 2023), https://openai.com/blog/data-partnerships.

[85] Colin Allen, *Artificial Morality: Top-down, bottom-up, and hybrid approaches*, 7 ETHICS & INFO TECH. 149, 149 (2005).

ethical agents in order to learn what behaviors are appropriate and which aren't through positive and negative reinforcements.[86]

Some scholars, in contrast, argue that Social Choice Theory should inform our training of AIs.[87] Social choice theory originated in the work of economist and political scientist Kenneth Arrow, and is a field of study that aims to develop methods for aggregating individual preferences or viewpoints into a collective decision or social welfare function.[88] The ultimate goal is to find ways to combine the diverse opinions and values of many individuals into a single, overarching viewpoint that represents the group as a whole.[89] However, asking *humans* to aggregate desires and values into a collective decision is fraught with difficulties.[90] Even the simplest method of majority voting can yield paradoxical results, with group preferences becoming fluid even when individual preferences remain consistent.[91] But AIs are uniquely positioned to employ social choice theory, and to do so in the context of filling the role of a jury.[92]

To illustrate this, consider how an AI jury might aggregate values across societies. Imagine two communities with different conceptions of fairness. Community X believes that fairness is best achieved by maximizing individual liberty. They prioritize personal freedom, property rights, and minimal government intervention. In their view, a fair society is one where people are free to pursue their own goals and reap the rewards of their efforts without undue restrictions.

In contrast, Community Y sees fairness through the lens of equality. They believe that a fair society is one that actively promotes equal opportunities and outcomes, even if that requires some limits on individual liberty. This might include

---

[86] Colin Allen, *Artificial Morality: Top-down, bottom-up, and hybrid approaches*, 7 ETHICS & INFO TECH. 149, 151 (2005).

[87] Seth D. Baum, *Social Choice Ethics in Artificial Intelligence*, 35(1) AI & SOC. 165 (2020), doi:10.1007/s00146-017-0760-1.

[88] *See generally* KENNETH ARROW, SOCIAL CHOICE AND INDIVIDUAL VALUES 1 (1951).

[89] *See generally* Amartya Sen, *The Possibility of Social Choice*, 89 AM. ECON. REV. 349 (1999).

[90] Christian List, *The theory of judgment aggregation: An introductory review* 187(1) SYNTHESE 179 (2012) (collecting research and explaining the "impossibility" of accurately aggregating human judgments.).

[91] This is known as Condorcet's paradox. Condorcet's paradox, also known as the voting paradox, occurs when collective preferences become intransitive (i.e., cyclic) even if individual preferences remain the same. For example, suppose there are three candidates A, B, and C, and three voters with the following preferences: Voter 1: A > B > C, Voter 2: B > C > A, and Voter 3: C > A > B. If the voters decide by majority rule, they will prefer A over B, B over C, and C over A, creating a cycle. This paradox demonstrates that majority voting can lead to inconsistent collective preferences, making it difficult to determine a clear winner. William V. Gehrlein, *Condorcet's Paradox*, 15(2) THEORY & DECISION 161 (1983). For other ways that people act contrary to their spoken interests, *see* Amos Tversky & Richard H. Thaler, *Anomalies: Preference Reversals*, 4 J. ECON. PERSP. 201, 202 (1990). Tversky and Thaler point out that, very often in economic and politics, people act in ways that are contradictory to their supposed preferences. "When people are asked to choose between a pair of options, a clear majority favors B over A. When asked to price these options, however, the overwhelming majority give values implying a preference for A over B." In addition, *see* Baruch Fischhoff, *Value Elicitation: Is there anything in there?* 46(8) AM. PSYCH. 835 (1991). Fischhoff discusses the challenges of eliciting people's true values and preferences, highlighting how different elicitation methods can lead to different results and how people's stated preferences can be influenced by various contextual factors.

[92] Seth D. Baum, *Social Choice Ethics in Artificial Intelligence*, 35(1) AI & SOC. 165 (2020), doi:10.1007/s00146-017-0760-1.

progressive taxation, affirmative action policies, and a robust social safety net to level the playing field.

Now imagine a human jury tasked with deciding a property dispute between two neighbors on the border of both communities. If a jury is sourced from Community X, the jury would likely place greater weight on factors like individual property rights and contractual agreements. They might rule in favor of the neighbor who can demonstrate the clearest legal claim to the disputed land, even if that leads to an uneven distribution of wealth or resources. But if the human jury is sourced from Community Y, the jury would be more inclined to consider the broader social implications of its ruling. They might factor in the relative economic positions of the two neighbors and issue a judgment aimed at producing a more equitable outcome, even if that means overriding certain individual property rights. And if the human jury is sourced from both Community X and Community Y, the case may result in a hung verdict because neither side may be able to come to an agreement on what is fair.

An AI jury, by contrast, could analyze those same facts in light of the collected writings on property law and political philosophy spanning both histories and cultures. The AI could be instructed to identify the varying ways in which ethicists and jurists from both communities have conceived of the nature and limits of property rights in contexts of social interdependence. The LLM could use its incredible wealth of knowledge about both communities—knowledge that the jury members do *not* possess about their own communities—to identify areas of common ground between the two communities' positions. Although they differ on the specific topic of property rights, perhaps the two cultures share a bedrock commitment to the idea that people should follow through on their word. By giving greater weight to what the communities *do* agree on, as opposed to focusing on what they *don't* agree on—as a human jury is likely to do—the AI could reframe the dispute in a way that makes consensus more achievable.

The AI could also be instructed to adjust its aggregation method to account for the *intensity* of each community's conviction on the specific issue at hand. If the members of Community X express a fanatically held belief in the sanctity of the individual property rights in this case, while the egalitarian sentiments of Community Y are more diffuse and qualified, the AI may give somewhat greater voice to the former in its ultimate verdict because of the greater intensity. This context-sensitive weighting would allow the AI to craft a decision that respects the deeply held values most salient to the dispute.

In sum, the goal of an AI jury is not to arrive at a lowest-common-denominator compromise between different moral viewpoints, but rather to construct a higher-order synthesis of views that does justice to the fundamental concerns *across* communities. This is a delicate balancing act, to be sure, but one that AI systems are uniquely positioned to attempt by virtue of their vast knowledge, analytical prowess, and freedom from the cognitive and cultural biases that can constrain human reasoning.

B. Juries as Judging Witness Character

It's important to acknowledge that one of the key roles of juries is to assess the credibility of witnesses who testify at trial. Jurors are tasked with observing the demeanor, consistency, and plausibility of witness testimony and determining how

much weight to give it in their deliberations.[93] This function relies heavily on human intuition, emotional intelligence, and the ability to read nonverbal cues, and is one of the greatest successes of the jury system.[94]

Except it isn't. Studies have shown that people are no better at determining when someone is lying than if they were doing it by chance, and *especially* when they are on a jury.[95] A meta-review of 206 studies found that humans have an accuracy rate of 54% for lie-detecting—hardly better than a coin toss.[96]

In contrast, some of the new technology-based lie detection systems claim accuracy rates of up to 90%.[97] While these may be overstated, more tested methods like polygraphs have been found in repeated studies to have accuracy rates between 85-90%.[98] If a court were to combine all of the various ways to use machines to detect

---

[93] *See* Fifth Circuit Criminal Jury Instructions 1.09 (1990) ("You alone are to determine the questions of credibility or truthfulness of the witnesses. In weighing the testimony of the witnesses, you may consider the witness's manner and demeanor on the witness stand . . ."); *United States v. Barnard*, 490 F.2d 907, 912 (9th Cir. 1973) ("The jury is the lie detector in the courtroom.").

[94] *See generally* George Fisher, *The Jury's Rise as Lie Detector*, 107 YALE L.J. 575 (1997).

[95] *See also* George Fisher, *The Jury's Rise as Lie Detector*, 107 YALE L.J. 575, 708 n.606 (1997).

[96] Amit Katwala, *The race to create a perfect lie detector – and the dangers of succeeding*, THE GUARDIAN (Sep. 5, 2019), https://www.theguardian.com/technology/2019/sep/05/the-race-to-create-a-perfect-lie-detector-and-the-dangers-of-succeeding. *See also* George Fisher, *The Jury's Rise as Lie Detector*, 107 YALE L.J. 575, 708 n.606 (1997) ("Another study employed nursing students as experimental liars. Researchers told the students that their capacity to lie effectively was an important part of being a good nurse. The researchers then asked the would-be nurses to watch a video screen and to describe what they saw on the video as pleasant ocean scenes. At some point, the image on the screen switched to one of horrible carnage, but the nurses were told to describe it as a flowery park. Later the researchers showed videotapes of the nursing students to experimental jurors to test whether they could tell when the students were telling the truth and when they were lying about the images on the screen. The study found that very few viewers of the videotapes 'did better than chance' at this task.").

[97] CONVERUS, *EyeDetect Can Now Be Used for Single Issue, Diagnostic Tests for Investigations* (Jan. 30, 2019), https://converus.com/press-releases/eyedetect-can-now-be-used-for-single-issue-diagnostic-tests-for-investigations/.

[98] *United States v. Scheffer*, 523 U.S. 303, 333 (1998). In his famous dissent, Justice Stevens cited several credible scientific experiments on polygraphs and wrote that "There are a host of studies that place the reliability of polygraph tests at 85% to 90%."

lies,[99] and pair this with an LLM's ability to synthesize vast amounts of data, AIs could potentially assess witness credibility far more accurately than human jurors ever could.

However, the question of whether AI can and should replace the human role in lie detection is a complex one that raises significant legal, ethical and practical concerns. As one author discusses, there are issues around the reliability of these technologies in real-world settings, the risk of machine bias, and the potential violation of witnesses' privacy and privilege against self-incrimination.[100] Unpacking all of these considerations is an in-depth topic that deserves its own focused analysis.

For the purposes of our current discussion on AI's potential to emulate the representative and deliberative functions of juries, it's sufficient to note that human assessments of witness credibility are demonstrably fallible,[101] while AI-based systems show promise in this domain even if they are not yet suitable for wholesale adoption. The broader question of whether and how AI lie detectors could be integrated into legal proceedings is an important one, but also one that risks sidetracking us from the specific aspects of jury decision-making that are most analogous to the knowledge-aggregating and perspective-synthesizing capabilities of large language models like ChatGPT.

In addition, it's worth noting that while this paper focuses on the potential for AI to replace juries, AI systems could also be well-suited to take on the roles of judges, and in particular appellate judges.[102] Judges, like juries, are expected to be impartial

---

[99] And there are quite a few ways for machines to detect lies. *See* Robert Bradshaw, *Deception and detection: the use of technology in assessing witness credibility*, 37 ARB. INT'L 707, 709 (2021) ("In recent years, a new generation of lie detectors has emerged claiming to use advances in computing and neuroscience to uncover deception with greater accuracy, notably:

- facial expression recognition builds on the work of psychologist Paul Ekman on ''micro-expressions'. Ekman found that subjects may betray their feelings through momentary, involuntary expressions, such as a brief look of panic before inventing a story, or satisfaction at having successfully passed off a lie;
- eye-tracking measures eye movements, pupil dilation, and blinking as telltale signs of lying. One programme, EyeDetect, claims up to 90 per cent accuracy in field studies;
- voice stress analysis monitors changes in tone in the subject's voice to identify statements when they are under stress and may, therefore, be lying;
- linguistic analysis differs in that, rather than looking for a physiological response, it focuses on the language used by the subject. For example, studies have found that liars are less likely to use first-person pronouns and more likely to ''distancing language'; and
- functional magnetic resonance imaging (fMRI) tracks the flow of oxygenated blood around the brain, showing activity in those areas. The theory is that different areas of the brain are active (and will light up in a fMRI scan) when lying than when telling the truth.").

[100] *See generally* Robert Bradshaw, *Deception and detection: the use of technology in assessing witness credibility*, 37 ARB. INT'L 707 (2021).

[101] Amit Katwala, *The race to create a perfect lie detector – and the dangers of succeeding*, THE GUARDIAN (Sep. 5, 2019), https://www.theguardian.com/technology/2019/sep/05/the-race-to-create-a-perfect-lie-detector-and-the-dangers-of-succeeding. *See also* George Fisher, *The Jury's Rise as Lie Detector*, 107 YALE L.J. 575, 708 n.606 (1997).

[102] Mark W. Klingensmith, *Let's Talk, ChatGPT: What Will the Judiciary's Future Look Like?*, 97 FLA. BAR J. 26, 26 (2023) ("Appellate issues are presented to the courts by written submission, usually through briefs or motions, identifying the specific issues on appeal. These issues are phrased in a manner to allow appellate judges to analyze them according to an established body of law. The relevant underlying facts have been "found" by the lower tribunal; the appellate court considers those established facts according to the applicable law. Or, the court is asked to interpret the meaning of words or phrases in a law to properly apply to a given set of facts. Under either scenario, a computer program like [ChatGPT] could be programmed to provide answers to such questions submitted to it.").

decision-makers who apply the law to the facts of a case. AI's ability to process vast amounts of legal information, identify relevant precedents, and make consistent, unbiased decisions could make it a valuable tool in judicial decision-making.

In fact, AI's potential to assist or replace judges may be even greater in the context of appellate courts, where the focus is on reviewing the application of the law rather than assessing the credibility of witnesses or weighing factual evidence.[103] Appellate judges often deal with complex legal issues that require sifting through large volumes of case law and legal scholarship, a task that AI systems are particularly well-equipped to handle.

However, the focus of this paper is on AI's potential to replace juries, given the similar functions of both. Juries are tasked with bringing the public's values, common sense, and collective judgment to bear on the facts of a case, a function that aligns closely with AI's ability to synthesize a wide range of perspectives and moral viewpoints from its training data.[104] While the potential for AI to replace judges is a topic worthy of further exploration, the unique characteristics of the jury's role as the community's conscience make it the primary focus of this paper's analysis.

IV. RESPONDING TO CRITICISMS OF AI IMPLEMENTATION IN THE LEGAL SYSTEM

A. What if the AI Makes a Mistake?

One of the most common critiques of using AI systems in high-stakes decision-making contexts is that we can't rely on their reasoning because *we don't know what their reasoning is*.[105] This is known as the "black box" phenomenon.[106] The most advanced deep learning models today, like GPT-4, operate in ways that are fundamentally inscrutable to outside observers—even to the AI developers themselves. One OpenAI coder said

> [W]e don't really know what [generative text AIs are] doing in any deep sense. If we open up ChatGPT or a system like it and look inside, you just see millions of numbers flipping around a few hundred times a second, and we just have no idea what any of it means. With only the tiniest of exceptions, we can't look inside these things and say, "Oh, here's what concepts it's using, here's what kind of rules of reasoning it's using. Here's what it does and doesn't know in any deep way." We

---

[103] Mark W. Klingensmith, *Let's Talk, ChatGPT: What Will the Judiciary's Future Look Like?*, 97 FLA. BAR J. 26, 26 (2023) ("Appellate issues are presented to the courts by written submission, usually through briefs or motions, identifying the specific issues on appeal. These issues are phrased in a manner to allow appellate judges to analyze them according to an established body of law. The relevant underlying facts have been "found" by the lower tribunal; the appellate court considers those established facts according to the applicable law. Or, the court is asked to interpret the meaning of words or phrases in a law to properly apply to a given set of facts. Under either scenario, a computer program like [ChatGPT] could be programmed to provide answers to such questions submitted to it.").

[104] *See supra* part II, III.

[105] Brent Daniel Mittelstadt et al., *The ethics of algorithms: Mapping the debate*, 3(2) BIG DATA & SOC. 1 (2016) ("Transparency is generally desired because algorithms that are poorly predictable or explainable are difficult to control, monitor and correct.").

[106] Davide Castelvecci, *Can we open the black box of AI?*, NATURE (Oct. 5, 2016), https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

> just don't understand what's going on here. We built it, we trained it, but we don't know what it's doing.[107]

As AI systems become more advanced and employ machine learning techniques like neural networks and genetic algorithms, the programmers and operators of these systems will increasingly lose their understanding over their exact decision making process.[108] Likewise, an AI jury's decision-making process would become inscrutable; we would be unable to determine the exact reason the AI had found the defendant guilty or innocent. In multiple ways, this could have undesirable consequences.

An AI jury may make a verdict decision that is clearly erroneous—like acquitting someone in the face of overwhelming evidence, or finding someone guilty despite scant evidence. If an AI jury were to acquit such a defendant, there may be no recourse for the government,[109] as it would be impossible to prove that the AI computed incorrectly or applied the wrong reasoning. But we must recognize that this is no different than a human jury.

Human juries very regularly return verdicts that are unexplainable in the face of the evidence, and yet we consistently uphold their verdicts as sacred.[110] We will even uphold a jury's decision when there were *substantial* defects in the deliberation

---

[107] Noam Hassenfeld, *Even the scientists who build AI can't tell you how it works*, VOX (Jul. 15, 2023), https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast.

[108] Andreas Matthias, *The Responsibility gap: Ascribing responsibility for the actions of learning automata*, 6 ETHICS INFO. TECH. 175, 182 (2004) ("In the course of the progression of programming techniques: from the conventional procedural program, via neural network simulations, to genetically evolved software, the programmer loses more and more of her control over the finished product. She increasingly becomes a 'creator' of 'software organisms', the exact coding of which she does not know and is unable to check for errors.").

[109] In the case of an acquittal, a defendant is constitutionally protected from being tried for the same crime again under the Double Jeopardy clause of the Constitution. U.S. CONST. amend. V; *see generally* Akhil Reed Amar, *Double Jeopardy Law Made Simple*, 106 YALE L. J. 1807 (1997).

[110] Two famous examples are the acquittals of O.J. Simpson and Casey Anthony. Despite compelling evidence, including DNA samples and blood stains linking Simpson to the crime scene, the jury found O.J. not guilty of the murders of Nicole Brown Simpson and Ronald Goldman. Many legal experts and observers believed that the verdict was influenced by factors such as the defense team's strategy of raising doubt about the handling of evidence and the racial composition of the jury. The Casey Anthony trial in 2011 resulted in another controversial jury verdict. Anthony was accused of murdering her two-year-old daughter, Caylee, and the prosecution presented evidence of Anthony's suspicious behavior and inconsistent statements. However, the jury found Anthony not guilty of first-degree murder, aggravated child abuse, and aggravated manslaughter of a child. The verdict shocked many who had followed the trial closely, as the evidence seemed to strongly suggest Anthony's involvement in her daughter's death. The jury's decision was seen by some as a failure of the justice system and a demonstration of the unpredictability of jury verdicts. *See generally* Christo Lassiter, *The O.J. Simpson Verdict: A Lesson in Black and White*, 1 MICH. J. RACE & L. 69 (1996); Nicholas A. Battaglia, *The Casey Anthony Trial and Wrongful Exonerations: How "Trial by Media" Cases Diminish Public Confidence in the Criminal Justice System*, 75 ALB. L. REV. 1579 (2011/2012).

process—like the jurors being drunk and high for the entire trial.[111] As Thom Brooks notes, "juries may convict or acquit for any reason acceptable to them" without needing to provide justification.[112] And we sanctify the voice of the jury in spite of the fact that human juries are *extremely* susceptible to biases, misunderstandings, and consideration of extra-legal factors.[113] The jury is—exactly like an AI—a "black box," where "meaningful review of the jury's decision-making process by appellate courts or the public is virtually impossible."[114]

To be sure, there are some methods of safeguards that allow us a peek into the jury's thought process—but these measure ultimately fall short of allowing for any meaningful remedies when biases or errors are uncovered.

One such safeguard is the use of special verdicts, where the jury is asked to answer a series of specific questions related to the case, rather than simply rendering a general verdict of guilty or not guilty.[115] Special verdicts can "improve the reliability of jury decision-making through the recognized psychological impact specific questions have in concentrating juror attention on certain matters to the exclusion of others."[116] By requiring jurors to focus on and respond to particular factual issues, special verdicts can potentially expose flaws in their reasoning that might otherwise remain hidden.[117]

However, special verdicts simply aren't employed. In civil cases, special verdicts "fail[ed] to gain wide acceptance" in courts and are "rarely used";[118] and in

---

[111] *Tanner v. US*, 483 U.S. 107, 121, 127 (1987). In *Tanner*, the Supreme Court upheld jury convictions for conspiracy and mail fraud even though there were allegations that several jurors consumed alcohol and drugs throughout the trial, causing them to sleep during the afternoons. The Court stated, "There is little doubt that postverdict investigation into juror misconduct would in some instances lead to the invalidation of verdicts reached after irresponsible or improper juror behavior . . . [Nevertheless,] long-recognized and very substantial concerns support the protection of jury deliberations from intrusive inquiry." *But see Pena-Rodriguez v. Colorado*, 137 S. Ct. 855, 855 (2017). In *Pena-Rodriguez*, the Supreme Court recognized a racial bias exception to the no-impeachment rule for jury deliberations. The Court held that where a juror makes a clear statement indicating that he or she relied on racial stereotypes or animus to convict a criminal defendant, the Sixth Amendment requires that the no-impeachment rule give way in order to permit the trial court to consider the evidence of the juror's statement and any resulting denial of the jury trial guarantee. The Court distinguished Pena-Rodriguez from Tanner, noting that racial bias, unlike the behavior in Tanner, implicates unique historical, constitutional, and institutional concerns and, if left unaddressed, would risk systemic injury to the administration of justice.

[112] Thom Brooks, *A Defence of Jury Nullification*, 10 RES PUBLICA 401, 402 (2004).

[113] *See generally* Lee J. Curley et al., *Cognitive and human factors in legal layperson decision making: Sources of bias in juror decision making*, 62(3) MED. SCI. L. 206 (2022).

[114] Kimberly A. Moore, *Judges, Juries, and Patent Cases – An Empirical Peek Inside the Black Box*, 99 MICH. L. REV. 365, 368 (2000) ("The 'black box' nature of jury verdicts leaves the Federal Circuit unable to correct inaccuracy or bias on the part of jurors."); Mark S. Brodin, *Accuracy, Efficiency, and Accountability in the Litigation Process – The Case for the Fact Verdict*, 59 U. CIN. L. REV. 15, 20 (1990).

[115] Mark S. Brodin, *Accuracy, Efficiency, and Accountability in the Litigation Process – The Case for the Fact Verdict*, 59 U. CIN. L. REV. 15, 20 (1990).

[116] Mark S. Brodin, *Accuracy, Efficiency, and Accountability in the Litigation Process – The Case for the Fact Verdict*, 59 U. CIN. L. REV. 15, 63 (1990).

[117] Mark S. Brodin, *Accuracy, Efficiency, and Accountability in the Litigation Process – The Case for the Fact Verdict*, 59 U. CIN. L. REV. 15, 20 (1990).

[118] Robert Dudnik, *Special Verdicts: Rule 49 of the Federal Rules of Civil Procedure*, 74 YALE L.J. 483, 488, 510 n.89 (1965).

criminal cases, "special verdicts are almost never used."[119] Without widespread adoption, special verdicts remain an insufficient tool for addressing the opacity of jury decision-making.

Another potential window into the jury's reasoning is post-verdict interviews with jurors. In high-profile cases, it's not uncommon for jurors to speak to the media about their deliberations and the factors that influenced their decision.[120] These types of interviews help shine a light into the "black box" of jury decision-making; but while they may satisfy public curiosity, they do little to ensure the fairness or accuracy of jury verdicts.

But these post-verdict interviews do nothing to fix errors that took place during the deliberations. The principle of finality protects jury verdicts from being impeached by testimony about what transpired during deliberations, except in very limited circumstances.[121] Federal Rule of Evidence 606(b) bars jurors from testifying about statements or incidents during deliberations, the effect of anything on their or other jurors' votes, or their mental processes in reaching the verdict.[122] So sure, we can peek into the jury-box after it has been opened, but this affords us no opportunity to fix whatever mistakes may have already been made by the jury.

In sum, while post-verdict interviews and special verdicts may offer glimpses into the jury's decision-making process, they ultimately provide little recourse for addressing any errors or biases that may have tainted the verdict. The jury remains a "black box," its inner workings largely shielded from meaningful scrutiny or correction—a characteristic shared by advanced AI systems like ChatGPT. The key difference between them, though, lies in our ability to proactively fix issues in reasoning *before* the jury starts to deliberate. We can address and mitigate biases in artificial intelligence far more easily than altering the deeply ingrained prejudices of human jurors.

B. The Systemic Biases in Jurors far Outweigh the Biases in AIs

The systemic biases that plague human jurors are far more severe and intractable than any biases found in AI systems. Implicit racial biases are pervasive among jurors, tainting their judgments in ways that are extremely difficult to detect—let alone correct.[123] Indeed, jurors themselves are often unaware of their own biases, and even when motivated to answer honestly, they may lack the self-knowledge to identify their

---

[119] Kate H. Nepveu, *Beyond "Guilty" or "Not Guilty": Giving Special Verdicts in Criminal Jury Trials*, 21 YALE L. & POL'Y REV. 263, 263 (2003).
[120] Nicole B. Casarez, *Examining the Evidence: Post-Verdict Interviews and the Jury System*, 25 HASTINGS COMM. & ENT. L.J. 499 (2003).
[121] FED. R. EVID. 606(b). *See also Tanner v. United States*, 483 U.S. 107, 107 (1987).
[122] FED R. EVID. 606(b).
[123] Anna Roberts, *(Re)forming the Jury: Detection and Disinfection of Implicit Juror Bias*, 44 CONN. L. REV. 827 (2012).

prejudices accurately.[124] And the effects of these biases are insidious, impacting not just ultimate verdicts but also the innumerable interpretations of evidence and witness credibility that occur throughout the trial. AI systems, in contrast, can be adjusted and optimized to reduce biases to negligible levels.

The easiest place to see this inherent bias in humans is when race is at play. Researchers have long found that white jurors are significantly more likely to convict a black defendant than a white defendant.[125] "[Even] when race was not a salient issue [in the trial,] White mock jurors did indeed demonstrate racial bias in their judgments. This racial bias could be seen not only in mock jurors' verdict and sentence recommendations, but also in their ratings of how strong the prosecution and defense cases were."[126]

Certainly, there are methods courts use to reduce the possibility of bias in jury members, the chief of these being *voir dire*. *Voir dire* is the process by which attorneys and judges question prospective jurors about themselves with the goal of identifying and removing those who may have biases or prejudices. Attorneys are permitted to ask

---

[124] Collin P. Wedel, *Twelve Angry (and Stereotyped) Jurors: How Courts Can Use Scientific Jury Selection to End Discriminatory Peremptory Challenges*, 7 STAN J.C.R. & C.L. 293, 310 (2011) ("[T]he overwhelming weight of evidence suggests that biased jurors are simply unaware of their biases."). Particularly relevant to this discussion of unfixable biases is the study of phenomenology. Phenomenology, a philosophical approach pioneered by Edmund Husserl, is the study of structures of consciousness as experienced from the first-person point of view. A key insight of phenomenology is that our experience of the world is always necessarily perspectival and situated. We encounter reality *exclusively* from a particular vantage point shaped by our unique personal histories, cultural contexts, and embodied experiences. This phenomenological understanding is highly relevant to the issue of juror bias. Jurors, like all humans, are inescapably confined to their own subjective lifeworlds—the pre-reflective, lived background that shapes their perceptions, interpretations and judgments. Their viewpoints are necessarily limited and conditioned by their individual horizons of experience. As such, jurors cannot help but approach a case through the lens of their own biases, prejudices and pre-understandings. Even with the best of intentions, they cannot completely step outside their own situatedness to arrive at a purely objective judgment. The phenomenological framework thus underscores the depth and intractability of juror bias, as it is rooted in the very structure of human subjectivity and the finitude of our experiential viewpoints. *See generally* EDMUND HUSSERL, IDEAS PERTAINING TO A PURE PHENOMENOLOGY AND TO A PHENOMENOLOGICAL PHILOSOPHY (F. Kersten trans., 1913); MAURICE MERLEAU-PONTY, PHENOMENOLOGY OF PERCEPTION (Colin Smith trans., 1945).

[125] Samuel R. Sommers & Phoebe C. Ellsworth, *WHITE JUROR BIAS; An Investigation of Prejudice Against Black Defendants in the American Courtroom*, 7 PSYCH., PUB. POL'Y, & L. 201, 202 (2001) ("Less than a century ago, anti-Black sentiment was accepted (and expected) among Whites, and the overtly prejudicial racial norms activated among jurors in racially-charged cases were not considered problematic."). This persistent racial bias in the criminal justice system is a central concern of Afro-Pessimism, a philosophical framework that posits that the western world is necessarily dependent on the continuation of anti-Black violence. As Frank B. Wilderson III argues in his influential book "Afro-Pessimism" (2020), the structural positioning of Black people as "socially dead" within the current global system renders them uniquely vulnerable to state violence and incarceration, and the white Western world depends on the perpetuation of this Black suffering to maintain its power and prosperity. The extraction of labor, the expropriation of resources, and the projection of white supremacist ideology all rely on the subjugated status of Black people. Efforts at reform or inclusion are thus viewed with skepticism, as they fail to address the fundamental antagonism between Blackness and the modern world order. Here, from an Afro-Pessimist perspective, the disproportionate conviction of Black defendants by white jurors is not merely a matter of individual bias but a manifestation of the underlying logic of anti-Blackness that pervades all aspects of social and political life, and one that is completely unfixable in the long-run. *See generally* FRANK B. WILDERSON III, AFROPESSIMISM (2020).

[126] Samuel R. Sommers & Phoebe C. Ellsworth, *WHITE JUROR BIAS; An Investigation of Prejudice Against Black Defendants in the American Courtroom*, 7 PSYCH., PUB. POL'Y, & L. 201, 220 (2001).

jurors about their backgrounds, opinions, and potential biases, and in some cases, jurors may be asked to fill out questionnaires about their views on relevant issues.[127] Based on the jurors' responses, attorneys can challenge jurors they believe will be unsympathetic to their side, either for cause (if the juror has a demonstrable bias or conflict of interest) or using a peremptory challenge (which requires no justification but is limited in number).[128]

However, in practice, *voir dire* is a highly imperfect tool for eliminating bias from juries. One major issue is that jurors may be hesitant to admit to biases, especially in a public setting in front of a judge, attorneys, and other jurors.[129] Implicit or unconscious biases are also particularly difficult to uncover through questioning because they operate on a subconscious level and may not be readily apparent even to the jurors themselves.[130] In sum, the process of *voir dire* is so ineffectual that one author wrote, "even the most extensive and penetrating *voir dire* will not screen the vast majority of bigoted jurors."[131]

But where the biases of jurors are not solvable, the biases of AIs are. We are much more capable of addressing and fixing the biases that perpetuate AIs than we are capable of fixing human biases—significant research has already been dedicated to reducing these prejudices.[132] Certainly, AIs were substantially biased in their infancy. One study from 2016 found that "[o]nly 20 percent of the people predicted to commit violent crimes [by the AI algorithm] actually went on to do so . . . The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants."[133]

---

[127] CLARENCE DARROW, VERDICTS OUT OF COURT 315 (Arthur Weinberg & Lila Weinberg eds., Ivan R. Dee 1989) (1963) ("[E]verything pertaining to the prospective juror needs to be questioned and weighed: his nationality, his business, religion, politics, social standing, family ties, friends, habits of life and thought[, and] the books and newspapers he likes and reads . . . . Involved in it all is . . . above all, his business associates, residence and origin.").

[128] *Edmonson v. Leesville Concrete Co.*, 500 U.S. 614, 620 (1991) ("Peremptory challenges are permitted only when the government, by statute or decisional law, deems it appropriate to allow parties to exclude a given number of persons who otherwise would satisfy the requirements for service on the petit jury.").

[129] *See McDonough Power Equip., Inc. v. Greenwood*, 464 U.S. 548, 553 n.3 (1984) (noting that a juror deliberately did not disclose that his brother had been in a similar type of injury to the plaintiff). *See also* Jennifer H. Case, *Satisfying the Appearance of Justice When a Juror's Intentional Nondisclosure of Material Information Comes to Light*, 35 U. MEM. L. REV. 315 (2005).

[130] Anna Roberts, *(Re)forming the Jury: Detection and Disinfection of Implicit Juror Bias*, 44 CONN. L. REV. 827 (2012).

[131] DERRICK A. BELL, RACE, RACISM, AND AMERICAN LAW 331 n.2 (6th ed. 2008).

[132] Alekh Agarwal, *A Reductions Approach to Fair Classifications*, PROCEEDINGS OF THE 35TH INT'L CONF. MACH. LEARNING 60 (2018); Faisal Kamiran & Toon Calders, *Data preprocessing techniques for classification without discrimination*, 33 KNOWLEDGE & INFO. SYS. 1 (2012); Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*, (Oct. 27, 2016), https://arxiv.org/abs/1610.08452; Biran Hu Zhang et al., *Mitigating Unwanted Biases with Adversarial Learning*, 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (February 2, 2018); https://doi.org/10.1145/3278721.3278779; L. Elisa Celis et al., *Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees*, CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY '19 https://dl.acm.org/doi/pdf/10.1145/3287560.3287586.

[133] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Thanks to a plethora of attention to this issue, these prejudices have been greatly reduced.[134] A recent study from 2020 found that an AI trained to predict recidivism in felons "achieved [an] accuracy of 89.8% and 90.4% for African Americans and Caucasians,"[135] and another examined the types of recidivist algorithms in use by agencies today and found that "the fairness criteria were approximately met for both interpretable models for blacks/whites and males/females—that is, the models were fair according to these criteria."[136]

In summary, while both human jurors and AI systems can exhibit biases, the prejudices of AI are far more tractable. Through focused research and development efforts, the biases in AI algorithms can be identified, quantified, and systematically reduced to acceptable levels. Human biases, on the other hand, are often subconscious, difficult to detect, and resistant to change. Thus, in the long run, AI has the potential to provide a fairer and more impartial assessment of evidence and guilt than human jurors. However, even if AI can be made to be statistically unbiased, there remain significant social and psychological barriers to the acceptance of AI juries.

V. WHAT COULD THE LONG-TERM EFFECTS OF AI JURIES BE?

As explained previously, one of the purposes of juries is to "insure [the] continued acceptance of the laws by all of the people."[137] Juries instill confidence in the system, assuring the average citizen that our legal system possesses one last bulwark against the tyrannical rule of the elite; but the idea of AIs deciding the guilt or innocence of defendants does not tend to instill a great deal of confidence in our system.[138]

This discomfort is deeply etched into our collective cultural psyche—from the coldly logical HAL 9000 in 2001: A Space Odyssey to the ruthlessly efficient robot police of RoboCop.[139] Instinctively, we foresee a loss of agency in the face of inscrutable technological forces beyond our control or understanding, sustained by a

---

[134] Alekh Agarwal, *A Reductions Approach to Fair Classifications*, PROCEEDINGS OF THE 35TH INT'L CONF. MACH. LEARNING 60 (2018); Faisal Kamiran & Toon Calders, *Data preprocessing techniques for classification without discrimination*, 33 KNOWLEDGE & INFO. SYS. 1 (2012); Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*, (Oct. 27, 2016), https://arxiv.org/abs/1610.08452; Biran Hu Zhang et al., *Mitigating Unwanted Biases with Adversarial Learning*, 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (February 2, 2018); https://doi.org/10.1145/3278721.3278779; L. Elisa Celis et al., *Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees*, CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY '19 https://dl.acm.org/doi/pdf/10.1145/3287560.3287586.

[135] Bhanu Jain et al., *Reducing Race-Based Bias and Increasing Recidivism Prediction Accuracy by using Past Criminal History Details*, 13TH PERVASIVE TECH. RELATED TO ASSISTIVE ENVIRONMENTS CONF. 409 (2020), https://dl.acm.org/doi/pdf/10.1145/3389189.3397990.

[136] Caroline Wang et al., *In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction* (2022), https://arxiv.org/pdf/2005.04176.pdf.

[137] *Powers v. Ohio*, 499 U.S. 400, 407 (1991).

[138] Letter From Thomas Jefferson to Thomas Paine, 11 July 1789, THE PAPERS OF THOMAS JEFFERSON, VOL. 15: MAR. 1789 TO 30 NOV. 1789, 266–270 (Julian P. Boyd ed., Princeton University Press, 1958) ("I consider trial by jury as the only anchor ever yet imagined by man, by which a government can be held to the principles of its constitution.").

[139] 2001: A SPACE ODYSSEY (Stanley Kubrick Productions 1968); ROBOCOP (Orion Pictures 1987). Other examples include: Harlan Ellison, *I have no mouth, and I must scream*, IF: WORLDS OF SCIENCE FICTION Mar. 1967; Isaac Asimov, *All the Troubles of the World*, SUPER-SCIENCE FICTION, Apr. 1959; E.M. Forster, *The Machine Stops*, 1909.

fear that AI juries would supplant human empathy with the sterile, pitiless calculations of an algorithm.[140]

This undermining of public confidence in the system by algorithmic AIs is warned of by John Danaher in his article, *Threat of Algocracy*.[141] Danaher believes that increasing reliance on algorithmic decision-making systems, which he says will result in a society ruled by an "algocracy,"[142] pose a significant threat to the legitimacy of the public decision-making processes.[143] The incomprehensibility of advanced algorithmic systems would make it impossible for ordinary citizens to meaningfully participate in and understand how the decisions that affect their lives are made.[144] This lack of understanding in the civic process would destroy the core democratic values of transparency, accountability and consent of the governed, and thus poses "a significant threat to the [political] legitimacy of such processes."[145] Danaher is wrong.

---

[140] DANIEL DINELLO, TECHNOPHOBIA!: SCIENCE FICTION VISIONS OF POSTHUMAN TECHNOLOGY (2006) ("Posthuman technology threatens to reengineer humanity into a new machinic species and extinguish the old one. Science fiction shows that this process will subvert human values like love and empathy, revealing that the intrinsic principles of these technologies fortify genetic discrimination, social fragmentation, totalitarianism, surveillance, environmental degradation, addiction, mind control, infection, and destruction."). *See also* FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI 124 (2020) ("A robot bailiff calls your case, and you step past a gated barrier to a chair and desk set out for defendants. The judge-avatar begins to speak: 'You have been found guilty of a critical mass of infractions of the law. Your car has gone at least five miles per hour over the speed limit ten times over the past two years. You downloaded three films illegally last year. You smoked marijuana at a party. According to our algorithmic sentencing, optimal deterrence for this mix of offenses is forty points deducted from your credit score, a fine of 5 percent of your assets, and your agreement to install a home camera, algorithmically monitored, for the next six months, to ensure you do not violate the law again. If you wish to file an appeal, please enter your grounds for appeal in the Swift Justice app just downloaded onto your phone. If not, please exit the door you came in through, and the kiosk will have further instructions.'").

[141] John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29(3) PHI. & TECH. 245 (2016). Other authors have also argued that the usage of AI in governmental decision making poses a threat to the legitimacy of those institutions. *See* Ludvig Beckman et al., *Artificial intelligence and democratic legitimacy. The problem of publicity in public authority*, AI & SOC (2022).

[142] "Algocracy" is a portmanteau of "algorithm" and the suffix "-ocracy," denoting a form of government or social organization. In an algocracy, algorithms and automated decision-making systems play a dominant role in shaping policy, allocating resources, and regulating human behavior.

[143] John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29(3) PHI. & TECH. 245, 245 (2016).

[144] John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29(3) PHI. & TECH. 245, 251 (2016).

[145] John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29(3) PHI. & TECH. 245, 245 (2016).

A. People Trust AI Reasoning on a Variety of Topics Even More than Their Own Judgment, and *Especially* When They Don't Understand the AI's Process

One study found that people trust AI's moral judgment roughly the same as a human's, and in fact trusted its capacity to reason significantly higher than a human's.[146] They found that people have "a higher capacity trust, overall trust, and reliance on AI experts, but have somewhat higher moral trust and higher responsibility ascription for human experts."[147] This suggests that people view AI as more capable at the reasoning and analysis involved in decision making, even if their trust in its moral judgment hasn't been fully realized yet.[148]

In fact, research indicates that not only are people willing to trust the reasoning of AIs, they actually *prefer* AI reasoning to the judgments of fellow humans.[149] Across a series of experiments, Logg, Minson, and Moore found that participants relied more heavily on advice when they thought it came from an AI system rather than a person, a phenomenon the researchers dubbed "algorithm appreciation."[150] Strikingly, this held true across a variety of different contexts—from visual estimation tasks to business and geopolitical forecasts to even matters of personal taste like predicting romantic attraction.[151] "[The participants] even showed a willingness to choose algorithmic advice over their own judgment."[152] People seem to have an intuitive faith in the superior reasoning capacity of AI, to the point that understanding less about how the algorithm arrives at its conclusions actually *increases* trust.

---

[146] Suzanne Tolmeijer et al., *Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making*, (CHI Conf. Hum. Factors Computing Sys., 2022), https://doi.org/10.1145/3491102. 3517732. *See also* Jaap J. Dijkstra et al., *Persuasiveness of Expert Systems*, 17 BEHAVIOUR & INFO. SYS. 155, 160–61 (1998) ("Given the same advice, subjects thought an [automated system] to be more objective and rational than a human adviser, especially when the expert system advice was given in a production rule format . . . Objectivity and rationality seem to be persuasive cues (beliefs) that can make users accept advice of expert systems without examining it.").

[147] Suzanne Tolmeijer et al., *Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making*, (CHI Conf. Hum. Factors Computing Sys., 2022), https://doi.org/10.1145/3491102. 3517732. *See also* Jaap J. Dijkstra et al., *Persuasiveness of Expert Systems*, 17 BEHAVIOUR & INFO. SYS. 155, 160–61 (1998) ("Given the same advice, subjects thought an [automated system] to be more objective and rational than a human adviser, especially when the expert system advice was given in a production rule format . . . Objectivity and rationality seem to be persuasive cues (beliefs) that can make users accept advice of expert systems without examining it.").

[148] [148] Suzanne Tolmeijer et al., *Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making*, (CHI Conf. Hum. Factors Computing Sys., 2022), https://doi.org/10.1145/3491102. 3517732. *See also* Jaap J. Dijkstra et al., *Persuasiveness of Expert Systems*, 17 BEHAVIOUR & INFO. SYS. 155, 160–61 (1998) ("Given the same advice, subjects thought an [automated system] to be more objective and rational than a human adviser, especially when the expert system advice was given in a production rule format . . . Objectivity and rationality seem to be persuasive cues (beliefs) that can make users accept advice of expert systems without examining it.").

[149] Jennifer M. Logg et al., *Algorithm appreciation: People prefer algorithmic to human judgment*, 151 ORG. BEHAV. HUM. DECISION PROCESSES 90 (2019).

[150] Jennifer M. Logg et al., *Algorithm appreciation: People prefer algorithmic to human judgment*, 151 ORG. BEHAV. HUM. DECISION PROCESSES 90, 90 (2019).

[151] Jennifer M. Logg et al., *Algorithm appreciation: People prefer algorithmic to human judgment*, 151 ORG. BEHAV. HUM. DECISION PROCESSES 90, 99 (2019).

[152] Jennifer M. Logg et al., *Algorithm appreciation: People prefer algorithmic to human judgment*, 151 ORG. BEHAV. HUM. DECISION PROCESSES 90, 99 (2019).

A study from 2020 stated exactly this: people actually trust AI systems more when they *don't* understand the AI's reasoning.[153] The researchers found that "providing more insights into how [a machine learning] system arrives at its decision can have a *negative* effect on trusting behavior."[154] It seems the less people understand about how the AI reached its conclusions, the more they are inclined to simply trust its decisions.[155]

This trust in AI systems isn't just theoretical—on the contrary, our reliance on AIs is already systemic. Gravett writes that the usage of algorithmic decision-making touches nearly every part of our daily lives: "the news articles we read, the movies we watch, the people we spend time with, whether we get searched in an airport security line, whether more police officers are deployed in our neighborhoods, and whether we are eligible for credit, healthcare, housing, education and employment opportunities, among a litany of other commercial and government decisions."[156]

In fact, we trust AI decision-making so much that we already use it in the criminal adjudication process. 11 states and 178 additional counties use recidivism-prediction technologies when determining the sentencing for someone convicted of a crime.[157] For years, scholars and practitioners have advocated for putting these AIs into our sentencing procedures. They have argued that automated risk-assessment systems are "more efficient, unbiased, and empirically-based" than humans alone;[158] that the systems prevent judges from "sentencing blindly";[159] that the systems "minimize both the rates and the length of incarceration for low-risk offenders, resulting in lower budgetary costs and reduced social harms."[160]

Whether or not these programs work is irrelevant; the point is that *we trust them*. We have welcomed AI decision-making systems into some of the most consequential areas of criminal justice, allowing algorithms to influence outcomes that profoundly impact people's lives and liberty. This casts serious doubt on Danaher's argument that

---

[153] Philipp Schmidt et al., *Transparency and trust in artificial intelligence systems*, J. DECISIONS SYS. (2020).
[154] Philipp Schmidt et al., *Transparency and trust in artificial intelligence systems*, J. DECISIONS SYS. 2 (2020).
[155] Philipp Schmidt et al., *Transparency and trust in artificial intelligence systems*, J. DECISIONS SYS. 2 (2020).
[156] William H. Gravett, *Judicial Decision-Making in the Age of Artificial Intelligence*, 58 LAW, GOVERNANCE & TECH. 281, 282 (2024). *See also* COUNCIL OF ECON. ADVISORS, EXEC. OFF. OF THE PRESIDENT, THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE FUTURE OF WORKFORCES IN THE EUROPEAN UNION AND THE UNITED STATES OF AMERICA (2022). (stating that the use of AI in hiring practices, ranging from crafting job descriptions to matching applicants with open positions to screening resumes and even conducting initial interviews via chatbots makes it "hard to think of a place in hiring where AI is not appearing." *See also* Yang Shen & Xiuwu Zhang, *The impact of artificial intelligence on employment: the role of virtual agglomeration*, 11 HUMAN. & SOC. SCI. COMM. 1 (2024); Brittany Kammerer, *Hired by a Robot: The Legal Implications of Artificial Intelligence Video Interviews and Advocating for Greater Protection of Job Applicants*, 107 IOWA L. REV. 817 (2022).
[157] *Where are risk assessments being used?*, MAPPING PRETRIAL INJUSTICE, https://pretrialrisk.com/national-landscape/where-are-prai-being-used/ (last visited May 1, 2024).
[158] Kelly Hannah-Moffat, *The Uncertainties of Risk Assessment Partiality, Transparency, and Just Decisions*, 27 FED. SENT'G REP. 244, 244 (2015).
[159] J.C. Olsen, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. REV. 1329, 1340 (2011).
[160] William H. Gravett, *Judicial Decision-Making in the Age of Artificial Intelligence*, 58 LAW, GOVERNANCE & TECH. 281, 283 (2024).

the opaque reasoning of advanced AI systems poses a threat to the legitimacy of algorithmic governance.[161]

Danaher's concern rests on the premise that an inability to meaningfully participate in and understand how the AI makes its decisions would erode public confidence. But the evidence suggests the opposite—people are not only comfortable with, but actively prefer AI decision-making, even (and perhaps especially) when the AI's reasoning is not fully transparent or comprehensible to the average person.

In light of this, Danaher's worry that an "algocracy" would destroy consent of the governed appears unfounded. The governed are increasingly consenting to algorithmic decision-making across a range of sensitive domains. The rise of AI juries, rather than sparking a crisis of confidence, could be accepted as a natural extension of this trend. While the sterile, pitiless AI of science fiction may haunt our cultural imagination, the reality is that people crave the efficiency and surety of an all-powerful machine.

B. What Are the Dangers of Embracing this New Techno-Judiciary?

This article advocates for the inclusion of AI into jury deliberations and decision-making. But what this article does *not* advocate for is the unchecked embrace of AI into our social institutions. As a precaution, this next section explores some of the dangers of blindly accepting this technology.

As our AIs continue to provide uncannily accurate and insightful answers to an ever-widening range of questions, people may start to view it as an omniscient oracle rather than a powerful but ultimately man-made tool. [162] They would witness the AI's ability to predict future events, solve complex problems, and provide guidance on personal and professional matters with a level of clarity and wisdom that seems

---

[161] John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29(3) PHI. & TECH. 245 (2016).

[162] Unbelievably, this might already be occurring. One company offers an AI service called Quid that provides "crucial foresight" to remain ahead of "consumer behaviors and market dynamic[] shift[s]." One advertiser for Quid stated that Quid can "[s]ee the future" and that it enables companies to "[k]now [their] customers better than they know themselves." QUID, https://www.quid.com/solutions/marketing (last visited Apr. 7, 2024). *See also* Amy Taubman, *The Oracle's Algorithm: Why AI-Powered Customer Insights Hold the Key to Business Success*, LINKEDIN (Feb. 21, 2024), https://www.linkedin.com/pulse/oracles-algorithm-why-ai-powered-customer-insights-hold-amy-taubman-lwpje/. In addition, one author from 2014 characterized the type of AI that interacts with its users via a question-and-answer system as an "oracle." He wrote, "An oracle is a question-answering system. It might accept questions in a natural language and present its answers as text . . . We would want the oracle to give truthful, non-manipulative answers and to otherwise limit its impact on the world. Applying a domesticity method, we might require that the oracle should use only designated resources to produce its answer. For example, we might stipulate that it should base its answer on a preloaded corpus of information, such as a stored snapshot of the Internet, and that it should use no more than a fixed number of computational steps. To avoid incentivizing the oracle to manipulate us into giving it easier questions—which would happen if we gave it the goal of maximizing its accuracy across all questions we will ask it—we could give it the goal of answering only one question and to terminate immediately upon delivering its answer. The question would be preloaded into its memory before the program is run. To ask a second question, we would reset the machine and run the same program with a different question preloaded in memory." NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 146 (2014). Bostrom, unknowingly, *exactly* described ChatGPT almost a decade before it was released, and characterized it precisely as an "oracle."

superhuman. This could lead to a dangerous form of AI pseudo-worship, where we ascribe infallibility and perfect insight to the system precisely *because* of its inscrutable processes and authoritative output.

This, of course, parallels the age-old human tendency to place undue faith in mystical processes that seem to possess intelligence or knowledge beyond our understanding.[163] From the Oracle of Delphi in ancient Greece to modern-day psychics and fortune-tellers, we have always been drawn to oracles, prophecies, and divine revelations even when—or especially because—the basis for their insights were unclear or unverifiable.[164] The difference with AIs, though, is that a belief that an AI is omniscient might be *justified*.

People might be right to believe the AI is omniscient because, in comparison to us, it is. AIs possess a level of knowledge and understanding that far surpasses that of any individual human—or even the collective wisdom of humanity.[165] AIs have access to an unimaginably vast database of information, including every book, article, and website ever written, as well as real-time data from countless sensors and devices around the world.[166] It would be able to process and analyze this information at speeds and scales that are simply impossible for the human mind.[167] And when this type of seemingly all-powerful oracle is put into the position of a jury, it would be almost impossible for a society to *not* begin to venerate it as the final arbiter of cognition.

If this is right, where would this take us epistemologically? If AI juries become so engrained into our cultural expectations, would we even *know* when an AI jury had rendered an incorrect verdict? Or would we assume that the AI had seen something, some facet of fairness hidden in the data, that we couldn't, and therefore the machine must be right? [168]

Surrendering this capacity for ethical judgment to an AI may cause us to lose confidence in our ability to independently assess right and wrong. We may begin to assume that any disagreement with the AI must be due to our own intellectual or moral

---

[163] *See* Justin L. Barrett, *Revelation and Cognitive Science: an invitation*, *in* THE OXFORD HANDBOOK OF DIVINE REVELATIONS 518 (Balázs M. Mezei et al eds., 2021) ("Furthermore, as finite information processors, human minds naturally and automatically fill in informational gaps to make coherent meaning from what they experience . . . As cognitive science of religion has shown, humans may find certain aspects of divinity and divine order relatively easy to understand and receive from interaction with the natural world."). Peter T. Struck also points out that people across the ancient Mediterranean world commonly assumed that "clandestine signs were buried in the world around them," and perceived meaningful messages in a wide array of natural phenomena. While moderns may view this as primitive superstition, Struck contends divination was the ancients' way of grappling with the human experience of "surplus knowledge"—the sense that "our ability to know exceeds our capacity to understand that ability." PETER T. STRUCK, DIVINATION AND HUMAN NATURE: A COGNITIVE HISTORY OF INTUITION IN CLASSICAL ANTIQUITY 3, 15 (2016).

[164] PETER T. STRUCK, DIVINATION AND HUMAN NATURE: A COGNITIVE HISTORY OF INTUITION IN CLASSICAL ANTIQUITY 3, 15 (2016).

[165] *Supra* part I.

[166] *Supra* part I.

[167] *Supra* part I.

[168] *See* Jaap J. Dijkstra et al., *Persuasiveness of Expert Systems*, 17 BEHAVIOUR & INFO. SYS. 155, 160–61 (1998) ("Given the same advice, subjects thought an [automated system] to be more objective and rational than a human adviser, especially when the expert system advice was given in a production rule format . . . Objectivity and rationality seem to be persuasive cues (beliefs) that can make users accept advice of expert systems without examining it.").

shortcomings. "That's what the machine says" could become a way to shrug off difficult moral quandaries or complex decisions. This might seem far-fetched, but consider this: in modern society, the definitive way to end an argument is to tell someone to "Google it."

One author argues that this moral outsourcing could become so extant that every action of our lives could become moderated by an AI "nannystate."

> Thanks to smartphones or Google Glass, we can now be pinged whenever we are about to do something stupid, unhealthy, or unsound. We wouldn't necessarily need to know why the action would be wrong: the system's algorithms do the moral calculus on their own. Citizens take on the role of information machines that feed the techno-bureaucratic complex with our data. And why wouldn't we, if we are promised slimmer waistlines, cleaner air, or longer (and safer) lives in return?[169]

In fact, this AI nannystate is actually *advocated for* by some parties. An article printed in *Corrections Today*, the journal for the National Institute of Justice, stated that

> [t]hrough wearable devices or smartphones, AI could reinforce programming with reminders, encouraging messages, and even warnings (depending on the mood and behavior of the individual) by monitoring the stress level of the individual or assessing the known attributes of the individual's physical location . . . With this new technology, jurisdictions can experiment with corrections reform while promoting successful reentry of more high-risk individuals.[170]

This should give us serious pause. While the allure of a jury system that fully and accurately reflects the morals of the community should be encouraged, we must consider the profound implications of surrendering our minute-to-minute autonomy to algorithmic overseers. The idea that AI could monitor our every move, thought, and feeling—issuing "warnings" whenever we step out of line—is a dystopian vision that would make even George Orwell shudder.

But perhaps the most insidious aspect of this AI-driven future is how readily we might accept it. As we grow increasingly accustomed to the convenience and surety of algorithmic decision-making, we may willingly trade our privacy and agency for the promise of a frictionless, optimized existence. We may come to see the AI's constant surveillance and behavioral corrections not as invasive and authoritarian, but as comforting and necessary—a benevolent guardian that knows what's best for us, even when we don't know it ourselves.

This is the true danger of an unchecked embrace of AI in our social institutions. It's not just about the loss of jobs or the erosion of human skills; it's about the slow, tectonic shift in our perception of what it means to be human. If we outsource our moral reasoning and decision-making to machines, we risk losing touch with the very qualities

---

[169] Evgeny Morozov, *The Real Privacy Problem*, MIT TECH. REV. (Oct. 22, 2013), https://www.technologyreview.com/2013/10/22/112778/the-real-privacy-problem/.

[170] Eric Martin & Angela Moore, *Tapping into artificial intelligence: Advanced technology to prevent crime and support reentry*, CORRECTIONS TODAY 28 (May/June 2020).

that define us as autonomous, thinking beings. We risk becoming passive subjects in a world governed by algorithms, our lives shaped by the inscrutable logic of an artificial intelligence that we no longer understand—or question.

CONCLUSION

As we stand at the cusp of a new era in legal decision-making, the potential for AI to revolutionize the jury system is both exhilarating and alarming. This article has argued that advanced language models like ChatGPT are uniquely positioned to serve as superior embodiments of the "community conscience" that juries are meant to represent. By synthesizing vast amounts of knowledge and diverse perspectives into nuanced, context-sensitive judgments, AI juries could do justice to the broader values and concerns of society in ways that 12-person human juries often fail to achieve.

The core of this argument rests on two key points: First, AI systems like ChatGPT are trained on enormous datasets spanning a vast range of human knowledge and viewpoints, effectively capturing something much closer to the total awareness and sensibilities of an entire society than any small group of individuals ever could. Second, the role of juries is to implement the community's moral and ethical standards when applying the law. By tapping into the written "conscience" of humanity writ large, AI is poised to fulfill this function more comprehensively and impartially than traditional juries.

However, this article has also explored the dangers of surrendering this moral agency. An overreliance on AI moral judgments could lead to a dangerous atrophying of human ethical reasoning and agency. If AI decisions come to be seen as infallible and unchallengeable, we risk creating a society where people no longer engage in the difficult—but necessary—work of grappling with moral quandaries and taking responsibility for their decisions. We may become passive subjects in a world governed by algorithms, our thought-processes shaped by the inscrutable logic of an artificial intelligence that we no longer question or understand.

In the end, the question is not whether AI will transform our legal system, but how we will shape that transformation. Will we succumb to the temptation of an algocratic nannystate, where every minute of our lives is governed by the inscrutable logic of machines? Or will we find a way to integrate AI into our judicial processes that honors the best of both human and machine intelligence? The answer lies in our willingness to grapple with these complex issues head-on, to think critically about the role we want technology to play in our society, and to assert our values in the face of an uncertain future. Only by engaging in this difficult but necessary conversation can we hope to build a justice system that is not only more efficient and accurate, but also more equitable, transparent, and accountable to the people it serves.

"DIGITAL PERSONAL DATA PROTECTION ACT" —A STRUDEL SERVED RAW!

Soumya Banerjee*

Abstract: The good news is that personal data privacy law has become a reality in India. The bad news is 95% of the public is ignorant of the same, 3% do not know what it entails and the rest 2% are voraciously engaged in an intellectual feud. Privacy as a concomitant of natural or inalienable rights has been recognized in the western hemisphere since the 18th century. In India, the legislative history of the doctrine can be traced to the trilogy of cases during the 1950's. It was only post 1978, that the apex court of the country passed a slew of judgements recognizing the right to privacy as an essential part of the right to "protection of life and personal liberty" embodied under the Indian Constitution and thereby bestowing the fundamental right status. On August 24, 2017, a nine-judge Constitution bench of the Supreme Court of India (Puttaswamy judgment) re-wrote history as it not only recognized and reconfirmed the fundamental right status of " right to privacy" but also laid the foundation of data protection law in the country. August 11, 2023, marks a historic date in the legislative annals of digital India, as the country enacted the Digital Personal Data Protection Act, 2023 ("DPDP Act" or "Act") after more than half a decade of deliberations. At a time when technology has become the defining paradigm of every business, the DPDP Act seeks to lay the foundation for developing a strong data privacy regime in the country. The Act in its new avatar is quite different from its predecessors proposed earlier. Ironically, the regulations themselves have set ajar a host of challenges, issues and steeplechases, which can barely be fathomed at this moment. In addition, the DPDP Act is yet to be notified or implemented by the Central Government. The key question this paper discusses is whether this seemingly endless period of deliberations culminated into a robust and comprehensive law or is it simply a Strudel served raw! To answer this question, the paper first charts the history of the concept with a chronological approach on a global platform. The second part of the paper charts the pre-DPDP era in India. The third part recapitulates the DPDP Act (in the present form) in a nutshell, while the fourth part dissects the DPDP Act highlighting certain potentially problematic features of this law. Lastly, the paper will examine what can be done to influence the development of a robust and sustainable data protection regime in the country in the years to come.

Keywords: Data Privacy; DPDP Act; Indian Privacy Law; Origin of Privacy Laws

* Legal Department, Yes Bank Limited, India.

Table of Contents

Introduction 88

I. Origins of the Concept Called "Privacy" 88

II. Privacy Law in India Prior to DPDP Act 91

III. Digital Personal Data Protection Act–in a Nutshell 93

A. Applicability 93

B. Data Processing Principles 94

C. Consent & Notice 94

D. Obligation of the Data Fiduciary 95

E. Significant Data Fiduciary 95

F. Processing of Personal Data of Children 96

G. Rights of the Data Principal 96

H. Cross-Border Transfer of Personal Data 96

I. Data Protection Board of India 97

J. Jurisdiction of the Board and Penalties 97

IV. A Half-Baked Strudel 97

A. Flawed Legislative Approach 97

B. Allied Laws 98

C. Applicability 99

D. Government—the Unregulated Hand 100

E. Data—the Modern Gold 100

F. Consent 101

G. Non-Consensual Processing 102

H. Data Fiduciary 102

I. Missing Ingredients of the Act 103

J. A Toothless Board 104

K. Dysfunctional Dispute Resolution Mechanism 104

L. Technologically Agnostic 104

V. Towards a Robust and Sustainable Framework 105

Conclusion 107

"You're the Apfelstrudel of mein eye"

–Chitty Chitty Bang Bang (1968)

INTRODUCTION

The good news is that personal data privacy law has become a reality in India. The bad news is 95% of the public is ignorant of the same, 3% do not know what it entails and the rest 2% are voraciously engaged in an intellectual feud. The latter two layers emerge primarily because of the different aspects involved in privacy, *viz., need, function, right, technology and legal protection.*[1] Just like a *Strudel* – a whirlpool of technique, process, and ingredients. For all those non aficionados of confectionary out there, *Strudel* is a type of sweet or savory layered pastry, where the filling is spread intermittently between layers of unleavened dough giving it a swirling pattern. Coincidentally, the oldest known strudel recipe[2] and the first ruling[3] by a Court of law recognizing the need for privacy (*though in a rudimentary form*), both find their origins in the 16th century.

August 11, 2023, marks a historic date in the legislative annals of digital India, as the country enacted the Digital Personal Data Protection Act, 2023 ("**DPDP Act"** or **"Act"**)[4] after more than half a decade of deliberations. At a time when technology has become the defining paradigm of every business, the DPDP Act seeks to lay the foundation for developing a strong data privacy regime in the country. Ironically, the regulations themselves have set ajar a host of challenges, issues and steeplechases, which can barely be fathomed at this moment. In addition, the DPDP Act is yet to be notified or implemented by the Central Government, just like a batch of freshly baked strudel resting in a rack.

The key question this paper discusses is whether this seemingly endless period of deliberations culminated into a robust and comprehensive law or is it simply a *Strudel served raw*! To answer this question, the paper first charts the history of the concept with a chronological approach on a global platform. The second part of the paper charts the pre-DPDP era in India. The third part recapitulates the DPDP Act (*in the present form*) in a nutshell, while the fourth part dissects the DPDP Act highlighting certain potentially problematic features of this law. Lastly, the paper will examine what can be done to influence the development of a robust and sustainable data protection regime in the country in the years to come.

I. ORIGINS OF THE CONCEPT CALLED "PRIVACY"

"Privacy" as a concept was alien to the early human civilization. The early homo sapiens, characterized by their bipedalism and subsistence lifestyle, largely lived in

---

[1] Karl de Leeuw and Jan Bergstra, (Eds), "The History of Information Security: A Comprehensive Handbook". Elsevier: 2007. Holvast, Jan, "History of Privacy", Holvast & Partner, Privacy Consultants, NL - Landsmeer, The Netherlands.

[2] The oldest Strudel recipe for a Millirahmstrudel, is from 1696, in a handwritten recipe at the Viennese City Library, Wiener Stadtbibliothek.

[3] *Semayne* v. *Gresham* (1604) 5 Co Rep 91; 77 ER 194 ('*Semayne's Case*').

[4] The Digital Personal Data Protection Act, 2023 (No. 22 of 2023), Gazette of India, August 11, 2023.

shared common dwellings with almost no physical separation. This meant no person could escape the physical surveillance of others without special efforts [5] and consequently resulted in little or no privacy. However, as humans transitioned from being gatherers to settlers in small encampments, the first seeds of privacy were sown.

Historically, the concept of privacy can be traced to the writings of Socrates, Plato and other Greek philosophers,[6] which noticeably distinguished between the 'outer' and 'inner', 'public' and 'private'; and 'society' and solitude'. From a legal perspective, the Code of Hammurabi contained a paragraph against the intrusion into someone's home.[7] Chronological research shows that 'physical privacy' was overtly recognized in England centuries ago. According to the Electronic Privacy Information Centre, the Justices of Peace Act of 1361 provided for the arrest of peeping toms and eaves dropper. [8] The concept of privacy formally took silhouette during the colonization of North America. Hence, when a large population of the early colonists migrated to the North America from England it was not surprising that the concept also sailed along with them and they started respecting privacy (during the early 16th century), in relation to an individuals' home, family and even correspondence. Ownership and possession of the land in the vast US continent laid the foundation for the privilege of privacy. Physical privacy became the characteristic of everyday life and home became the primary place of privacy.[9] Reverberation of the same cogitate can also be seen in the famous ruling by Sir Edward Coke in the *Semayne Case [January 1604]*[10]-"*a man's home is his castle.*"

The 19th century witnessed a new series of threats which fuelled the rise of progressive regulations in the field of privacy. Preventing-copies of US census being published in public [11], unauthorized opening of mail [12], tapping into telegraph communication[13], compelling disclosure of personal information and documents[14], are some of the prominent cases in the US legal history. Then came the Warren and Brandeis publication in 1890[15], which is widely recognized as the cradle for the concept and came to be considered as the "most influential law review article of all"[16] for more than one reason. First, it highlighted the role of media [newspaper/prints] which transgressed from a mere information source to indulge in "yellow journalism". Second, the lack of common law remedy available at that time to combat such threat and lastly,

---

[5] David H. Flaherty, Privacy in Colonial New England, 2 (1972).
[6] Moore Jr., B.: Studies in Social and Cultural History. M.E. Sharpe, Inc., Armonk (1984).
[7] Solove, D. J.: *Nothing to Hide: The False Tradeoff between Privacy and Security*, New Haven & London: Yale University Press, 2011. p. 4.; Lukács Adrienn: *What is Privacy? The History and Definition of Privacy*, Országos Szövetsége, Budapest, Magyarország, pp. 256-265. (2016).
[8] Holvast Jan, *History of Privacy*, Karl de Leeuw and Jan Bergstra (Eds), The History of Information Security: A Comprehensive Handbook. Elsevier, (2007).
[9] Flaherty, D.H.: Privacy in Colonial New England. University Press of Virginia, Charlottesville (1972).
[10] Court of King's Bench, All ER Rep 62, Also reported 5 Co Rep 91 a; Cro Eliz 908; Moore KB 668; Yelv 29; 77 ER 194. *Supra Note 3.*
[11] David J. Seipp, The Right to Privacy in American History, 6–7 (1978).
[12] Robert Ellis Smith, Ben Franklin's Web Site: Privacy and Curiosity from Plymouth Rock to the Internet, 12 (2000).
[13] *Ex parte* Brown, 72 Mo. 83, 95 (1880). See *Supra Note* 16 at 5.
[14] *Boyd* v. *United States*, U.S. Supreme Court, 616 (1886).
[15] Samuel D. Warren & Louis D. Brandeis, The Right to Privacy, 4 HARV. L.REV. 193 (1890).
[16] Harry Kalven, Jr., Privacy in Tort Law—*Were Warren and Brandeis Wrong*?, 31 L. & CONTEMP. PROBS. 326 (1966).

it advocated numerous remedies, with the principal being "an action of tort for damages in all cases"[17], for the preservation of privacy. The publication eventually led to State of New York enacting a statute[18] establishing a cause of action for invasion of privacy and subsequent introduction of statutes to safeguard against unauthorized intercepting of telegraph and telephone calls.[19]

With the dawn of the 20th century, several statutes were enacted across the world (both federal and state) and Courts vigorously recognized, reiterated, and ruled in favor of protection of privacy. Numerous statutes[20] were enacted between 1960-1980, which dealt with safeguarding privacy of individuals or their information and influenced privacy laws across the globe. The German State of Hessian[21] enacted the World's first data privacy laws at State level in September 1970, and subsequently, laid the foundation for the German Constitutional Court to recognize the fundamental right of informational self-determination[22]. In 1973, Sweeden enacted the Data Act, which is one of the first privacy laws related to computers and online activities. The next ten years in US saw rapid enactment of statutes, which aimed safeguarding the privacy of individuals against unauthorized collection[23], dissemination[24] and usage[25] of personal information *via* various communication channels.

At the international level, the United Nation ("UN") Declaration of Human Rights[26] enshrines the right to privacy under Article 12. In 1950, the European Convention of Human Rights reiterated similar protection under Article 8, subject to certain restrictions. The Organization of Economic Cooperation and Development (OECD) Privacy Guidelines [Eight Principles][27] in 1980, charts the formal birth of information privacy laws at international level. However, it was Convention 108 in 1981[28], the first binding international instrument, which aimed at protecting individuals against abuses derived from the collection and processing of personal data and sought to regulate the cross-border flow of personal data. Thereafter in 1996, the European Union promulgated the Data Protection Directive,[29] which established the basic principles for privacy legislation for EU member states and provided for a comprehensive protection of personal information, including restricting the flow of personal data outside the borders of EU. This broad-brush approach was a stark contrast to the United States' approach, which regulated privacy at a "sectoral level" in various

---

17 *Supra Note 23, 219.*
18 New York Civil Rights Law, 50–51.
19 *Supra Note 17, 157.*
20 Freedom of Information Act of 1966. Code of Fair Information Practices of 1973. Marc Rotenberg, Fair Information Practices and the Architecture of Privacy (What Larry Doesn't Get), 2001 STAN. TECH. L. REV. 1, 44. Privacy Act of 1974. Foreign Intelligence Surveillance Act of 1978. Fair Credit Reporting Act of 1970. Bank Secrecy Act of 1970.
21 Data Protection Act, 1970; Datenschutzgesetz [HE 1970]. GVBl. HE 1970 S. 625.
22 Bundesverfassungsgericht. Judgement of the first senate of 15. December 1983.
23 Cable Communications Policy Act of 1984.
24 Privacy Protection Act of 1980; Video Privacy Protection Act of 1988.
25 Computer Matching and Privacy Protection Act of 1988.
26 Declaration was proclaimed by the United Nations General Assembly in Paris on 10 December 1948 (Resolution No.217A).
27 Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data of 23 September 1980 C (80)58/FINAL.
28 The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, January 1981.
29 Council Directive 95/46, 1995 O.J. (L 281) 31–50 (EC), hereinafter "EU Data Protection Directive."

narrow contexts.[30] In 2014, the African continent adopted the Malabo Convention[31] which established the legal framework for personal data protection and cyber security with the African Union Member States, along with the mechanism for combating violation of privacy in relation to data collections, processing, transmission, storage and usage.

Between 2016-19, data privacy regulation received a big push, with the introduction of GDPR[32] and ePrivacy Regulation[33] in the European Union. GDPR became a global sensation on account of three reasons: (i) "*Brussels Effect*"[34], because of its aggressive extraterritorial scope and imposition of EU laws across the world. (ii) "*DC Effect*"[35], because of its adoption of various US data privacy innovations, e.g., privacy by design, deterrence-based fines, corporate fines and compensation from law-breaking date processors; and (iii) the "*Individual focused approach*", because of the elaborate rights it gave to individuals, e.g., right to be forgotten, object, rectifications, portability, access and notifications. In 2020, the State of California became the first US State to enact a comprehensive data privacy law, which provided certain rights to customers and paved the way for other legislation in the US (*state and federal level)*. These days, January 28 of every year is celebrated as the 'Data Privacy Day' to commemorate the date when Convention 108 was opened for signature.

II. PRIVACY LAW IN INDIA PRIOR TO DPDP ACT

Privacy as a concomitant of natural or inalienable right[36] was recognized in the western hemisphere since the 18th century. In India, the legislative history of the doctrine can be traced to the trilogy of cases (M.P. Sharma[37] - Kharak Singh[38] - Gopalan[39]) during the 1950's. Ironically, the first two judgements refused to recognize the right of privacy as a fundamental right, *in absentia* of express provisions in the Indian Constitution, while the third judgment simply assumed the existence of such rights emanating from personal liberty while subjecting it to restriction on the basis compelling public interest (based on the state test under US jurisprudence).

It was only post 1978, that the apex court of the country passed a slew of judgements recognizing the right to privacy as an essential part of the right to "protection of life and personal liberty" embodied under the Indian Constitution and

---

[30] Joel R. Reidenberg, Setting Standards for Fair Information Practices in the U.S. Private Sector, 80 IOWA L. REV. 497 (1995). Daniel J. Solove, A Brief History of Information Privacy Law, Chapter-1, 1.4.4 (2006).

[31] The African Union Convention on Cyber Security and Personal Data Protection

[32] General Data Protection Regulation, Regulation (EU) 2016/679.

[33] Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications, 2018.

[34] Anu Bradford, The Brussels Effect, 107 NW. U. L. REV. 1, 1 (2012).

[35] Michael L. Rustad and Thomas H. Koenig, Towards a Global Data Privacy Standard, 71 Fla. L. Rev. 365.

[36] American Declaration of Independence (1776); Declaration of the Rights of Man and of the Citizen (1789).

[37] *M.P. Sharma and Others* v. *Satish Chandra, District Magistrate, Delhi and Others*, 954 AIR 300, 1954 SCR 1077, AIR 1954 Supreme Court 300, 56 PUN LR 366.

[38] *Kharak Singh* v. *State of U.P. And Others*, 963 AIR 1295, 1964 SCR (1) 332, AIR 1963 Supreme Court 1295, 1963 ALL. L. J. 711, 1963 (2) CRI. L. J. 329, 1964 (1) SCR 332 1964 2 SCJ 107, 1964 2 SCJ 107.

[39] *A.K. Gopalan v. State of Madras, Union of India,* 1950 AIR 27, 1950 SCR 88, AIR 1950 Supreme Court 27, 1963 MADLW 638.

thereby bestowing the fundamental right status. The prominent cases involve impounding of passports (Maneka Gandhi[40]), telephone tapping (PUCL[41]), restrain on publication of material of a death row convict (Rajagopal[42]), inspection and search of confidential information (Canara Bank[43]), disclosure of HIV status of a patient (Mr. X v. Hospital Z[44]), medical termination of pregnancy (Suchita Srivastava[45]) and right of transgenders (NALSA[46]).

On August 24, 2017, a nine-judge Constitution bench of the Supreme Court of India (*Puttaswamy[47] judgment*) re-wrote history. The judgement not only recognized and reconfirmed the fundamental right status and reinforced the propositions laid down by above-mentioned judgments but also explicitly observed[48] that: (i) privacy is a constitutional core of human dignity; (ii) privacy safeguards personal autonomy and heterogeneity (iii) constitution must evolve with the felt necessities of time to meet the challenges thrown up in a democratic order governed by the rule of law (iv) any law which encroaches upon privacy will have to withstand the touchstone of permissible restrictions on fundamental rights and (v) informational privacy is a facet of the right of privacy.

Prior to the DPDP Act, there was no specific legislation on privacy and data protection in the country. Certain regulatory bodies (Telecom Regulatory Authority of India, Reserve Bank of India, Medical Council of India and Insurance Regulatory and Development Authority of India) under sector specific statutes[49] attempted to safeguard the interest of individuals by imposing restrictions on disclosure of information or documents to third parties, unless the same was required by law or the process prescribed therein. Nearly twenty-three years ago, the Information Technology Act, 2000[50] ("**IT Act**"), encapsulated provisions to protect the rights of individuals against breach of privacy by corporate entities. The IT Act was inspired by the Model Law on Electronic Commerce[51] encapsulated three elements of data protection, *viz.,* maintaining reasonable security practices and procedures to safeguard specified information classified as sensitive personal data or information which can identify a natural person ("**SPDI**"); recognition of tort remedy[52] upon breach in maintaining reasonable security practices and procedures, and lastly, the intentional disclosure of personal or sensitive information of any person, collected under a contractual relationship. The IT Act also attempted to protect the right of an individual against any unauthorized capturing, publishing, and transmission of any image of a private part of

---

40 *Maneka Gandhi* v. *Union of India* (1978) 1 SCC 248.
41 *People's Union for Civil* Liberties v. Union of India, (1997) 1 SCC 301.
42 *R. Rajagopal* v. *State of Tamil Nadu*, [(1994) 6 SCC 632; AIR 1995 SC 264.
43 *District Registrar and Collector, Hyderabad* v. *Canara Bank,* (2005) 1 SCC 496.
44 (1998) 8 SCC 296.
45 *Suchita Srivastava* v. *Chandigarh Administration*, (2009) 9 SCC 1.
46 *National Legal Services Authority* v. *Union of India*, (2014) 5 SCC 438.
47 *Justice K.S. Puttaswamy (Retd.) and Another* v. *Union of India*. (2019) 1 SCC 1.
48 *Ibid 262-265*.
49 Indian Telegraph Act, 1885; Banking Companies Act [Transfer and Acquisition of Undertakings], 1980; Credit Information Companies (Regulation) Act, 2005; Indian Medical Council Regulation, 2002; Insurance Regulatory and Development Authority of India, Regulation 2015 and 2017.
50 Act No.21 of 2000;
51 UNICITRAL Model Law on Electronic Commerce (June 12, 1996) with additional Article 5 bis as adopted in 1998.
52 Section 43A [Compensation for failure to protect data]; Ins. by Act 10 of 2009 (w.e.f. 27-10-2009).

such individual, under circumstances which violated his/her privacy[53], and thereby recognizing the sanctity of the body. In 2011, the Government introduced the Indian Data Protection Rules, which read with Section 43A of the IT Act, laid down eight rules to protect the privacy of an individual.

In July 2017, post the *Puttaswamy*[54] judgment, the Ministry of Electronics and Information Technology set-up the Srikrishna Committee[55], chaired by Justice B.N. Srikrishna (a retired judge of Supreme Court of India) to formulate the foundation of data protection norms in the country. The work of the committee formed the pedestal for the Personal Date Protection Bill of 2019[56], the first government version of the law. Unfortunately, the work of the committee and the resultant Bill of 2019 was more like saffron, white truffles and wagyu all rolled into one utopian savory strudel. Though the committee/Bill adopted a normative approach rather than US *laissez-faire* approach or the *individual dignity* centric approach adopted by EU, its expansive scope was hugely problematic, suffered from overregulation and the implementation framework more disruptive rather than transformative for the digital Indian economy. Nonetheless, it was perhaps the most comprehensive, cross-sectoral framework based on preventive requirements of business (known as "data fiduciaries") and right for individuals (known as "data principles").[57] The Bill of 2019 was withdrawn in November 2022, basis the report submitted by the Joint Parliamentary Committee, and replaced with the DPDP Bill of 2022[58]. The DPDP Bill of 2022 received the approval of the lower house (Lok Sabha) and the upper house (Rajya Sabha) in the month of August 2023 and officially became an Act after receiving the President's assent on August 11, 2023.

III. DIGITAL PERSONAL DATA PROTECTION ACT—IN A NUTSHELL

The DPDP Act comprises of 9 chapters, encompassing 44 sections and Schedule. In this part, only the key provisions of the Act have been enumerated rather than providing a narrative on the entire Act, which would have made a good case for writing a book but will surreptitiously defeat the scope and objective of this Article.

A. Applicability

The DPDP Act only applies to personal data collected from individuals, *i.e., Data Principal*[59] *in India* in a digital form or in non-digital form and digitized subsequently.[60] The Act applies to all data collected within India (*Territorial scope*) and processing of data outside the territory of India (*Extra-territorial scope*), if such processing is in connection with any activity related to offering of goods and services to Data Principal within the territory.[61] What this implies is- the Act is applicable to all individuals, who are Indian citizens, non-resident Indians and foreign citizens, if the

---

[53] *Supra Note* 60. Section 66E.
[54] *Supra Note* 58.
[55] *A Free and Fair Digital Economy, Protecting Privacy and Empowering Indians*, Committee of Experts under the Chairmanship of Justice B.N. Srikrishna, (2017).
[56] Bill No.373 of 2019.
[57] Anirudh Burman, *Understanding India's new Data Protection Law*, Carnegie India - Carnegie Endowment for International Peace, October 3, 2023; https://carnegieindia.org/2023/10/03/understanding-india-s-new-data-protection-law-pub-90624.
[58] Bill No.113 of 2022.
[59] As defined in *Section 2(j)*.
[60] *Section 3(a)*.
[61] *Section 3(b)*.

data is collected in India. This extra-territorial application and scope is an unique feature of the statute (similar to the PDPA[62] of Singapore), perhaps the first of its kind in the country, which explicitly includes all individuals and permits cross-border transfer of personal data to facilitate various e-commerce websites/international businesses operating and providing goods and services in India. The Act, however, excludes personal data processed by an individual for any personal or domestic purpose or made publicly available by the Data Principal or any person who is under an obligation under any law.[63]

B. Data Processing Principles

The DPDP Act lays down the *four elements*[64] for processing any personal data, *viz*., in accordance with the provision of the Act, lawful purpose, consent of the Data Principals, and for certain legitimate uses. The terminology - '*legitimate uses*' has rechristened the concept of "*deemed consent*", which was envisaged in the draft Bill of 2022, for processing of personal data for certain special use cases without the express consent of the Data Principal. The Act, similar to the provisions under GDPR and LGPD[65], list downs the legitimates uses[66], e.g., interest of sovereignty, integrity and security of India, fulfilling the obligations under law, responding to medical emergencies involving a threat to life or health, medical treatment, occurrence of disaster and purposes of employment. What this implies is the Data Principals will not have any right to erase, correct, access their personal data, or withdraw their consent for the original purposes for which it was disclosed. At the same time, the concept will reduce the dependency of obtaining express consent in specific circumstances, and ultimately result in cost savings[67] for the businesses due to the dispensation of additional mechanisms for consent management.

C. Consent & Notice

A valid consent under the DPDP Act needs to be free, specific, informed unconditional and unambiguous in nature with a clear affirmative action.[68] What this implies is – (i) consent cannot be obtained from Data Principles on a 'deemed'[69], 'omnibus' or 'conditional' basis; (ii) consent obtained for purpose A cannot be used for purpose B, and (iii) processing of such information shall be limited only to the personal data which is necessary for the specific purposes. To simply put it, it's an opt-in model of obtaining consent *akin* to GDPR and LGPD[70]. The Act, however, does not describe the form or manner of obtaining the consent through the electronic medium, like clickwrap, two-factor authentication etc. Nonetheless, any consent provided by the Data Principal shall not be absolute or permanent. It may be withdrawn at any time,

---

62 Singapore's Personal Data Protection Act, 2021.
63 *Section 3(c) (i)(ii)*.
64 *Section 4(1), (2)*.
65 Brazilian General Data Protection Law of 2020.
66 *Section 7*.
67 Decoding the Digital Personal Data Protection Act, 2023, KPMG India, August 2023.
68 *Section 6(1)*.
69 Unless the same is exempted under *Section 7*.
70 Brazilian General Data Protection Law of 2020.

either by the Data Principal or through its Consent Manager[71] and the ease of doing so needs to be at par with the standards adopted at the time of obtaining it.

Any notice sent to the Data Principal for the purpose of obtaining consent shall specifically inform such Data Principal about the personal data, purpose for which it is processed, manner of exercising their rights and making a complaint to the Data Protection Board[72]. Every request for consent shall be presented to the Data Principal in clear and plain language, with an option to access such request either in English or any language specified in the Eighth Schedule to the Constitution, including the details of the Data Fiduciary[73] or the Data Protection Officer ("**DPO**")[74], if applicable.[75]

D. Obligation of the Data Fiduciary

The Data Fiduciary, irrespective of any agreement to the contrary or failure by the Data Principal, is solely responsible and/or liable for all compliances under the Act and Rules made thereunder, including any processing of data undertaken by itself or any Data Processor[76] on its behalf. From implementing appropriate technical measures to taking reasonable security safeguards against any breach, Data Fiduciary is the phyllo dough of the Act. Any personal data being processed by the Data Fiduciary must ensure its completeness, accuracy and consistency.[77] They are also required to erase personal data if the specified purposes is served or if the Data Principal withdraws her consent, unless the retention of such data is mandated by law.[78] Additionally, they are also responsible for establishing effective mechanism to redress the grievances of the Data Principal.[79]

E. Significant Data Fiduciary

The Act defines a significant data fiduciary ("**SDF**") as any data fiduciary or class of data fiduciaries which is notified by the Central Government basis certain factors[80] enumerated under the Act, probably on account of the following three reasons: (i) to supervise the large regulatory space intersecting numerous business organizations across diverse sectors; (ii) establishing a supervisory regime for entities of national interest and 'too big to fail'; (iii) subjecting such entities to incremental compliance requirements such as – appointment of an individual as a data protection officer ("**DPO**") based in India, appointing an independent data auditor for evaluating compliance with the Act, conducting periodic audit and data protection impact assessment, and undertaking such other measures consistent with the provisions of the Act or as may be prescribed by the Central Govt., from time to time.

---

71 As defined in *Section 2(g)*.
72 *Section 5*.
73 As defined in *Section 2(j)*.
74 As defined in *Section 2(l)*.
75 *Section 6(3)*.
76 As defined in *Section 2(k)*.
77 *Section 8(3)*.
78 *Section 7(a)*.
79 *Section 7(10)*.
80 *Section 10(1)*.

F. Processing of Personal Data of Children

The Act expressly mandates that before processing any personal data of a child or a person with disability, it shall be mandatory for the Data Fiduciary to obtain a 'verifiable consent' of the parent or lawful guardian.[81] However, the Act explicitly forbids tracking or behavioral monitoring of, and targeted advertising directed at, children or a person with disability, and processing of children's data that is likely to cause any detrimental effect upon the child.[82] Notably, the Act empowers the Central Government to exempt certain classes of data fiduciaries and processing for certain purposes from the requirement of obtaining parental consent and prohibiting behavioral monitoring. It also empowers the Central Government to exempt data fiduciaries for processing data of children above a certain age but under 18 years in certain situations without the specific obligations attached to processing children's data. What this implies is - Data Fiduciaries need to implement suitable internal mechanisms for the purpose of obtaining and collecting 'verifiable' age of the child and consent of the parent/guardian, to safeguard against any detrimental effect upon such specific class of Data Principal. Though the Act seeks to provide enhanced safeguards for the vulnerable class, the proviso clearly lacks clarity in terms of what tantamount to a 'verifiable' consent or detrimental effect.

G. Rights of the Data Principal

Chapter III of the Act enumerates certain rights of the Data Principal, which includes the right to access information about personal data[83], right to correction and erasure[84], right to grievance[85] and the right to nominate[86].Of all the rights mentioned above, the last right, *i.e.*, to nominate, assumes special significance on account of three reasons: (i) it is a unique feature of law unparallel with any privacy laws across the world which allows the Data Principal to nominate any individual in the event of death or incapacity; (ii) recognizes personal data as an perceptible and inalienable property of an individual; (iii) allows the individual to control the personal information through a nominee rather than being freely available in the public domain.

H. Cross-Border Transfer of Personal Data

Like many data privacy regulations across the world (GDPR, PIPL[87] and nFADP[88]), the DPDP Act allows free transfer of data outside the territory of India for the purpose of processing.[89] However, such transfer is subject to notification by the Central Govt. in relation the country where data may or may not be transferred.

---

[81] *Section 9*.
[82] *Section 9 (3)*.
[83] *Section 11*.
[84] *Section 12*.
[85] *Section 13*.
[86] *Section 14*.
[87] China's Personal Information Protection law, 2021.
[88] New Federal Act on Data Protection of Switzerland, 2023.
[89] *Chapter IV Section 16*.

I. Data Protection Board of India

Chapter V of the Act contemplates the establishment of a Data Protection Board ("**Board**"), a body corporate having perpetual succession and a common seal under the aegis of the Central Government. The Board is slated to be an independent body and function as a digital office with receipt of complaints, hearing, pronouncement of decision impose of penalties and adopt such techno-legal measure as may be prescribed.[90] Any appeal preferred against an order of the DPB will be required to made before the Telecom Disputes Settlement and Appellate Tribunal ("**TDSAT**") established under the Telecom Regulatory Authority of India Act, 1997. Any appeal against the order of the TDSAT will be preferred before the Supreme Court of India. Two significant provisions deserve special attention here- (i) Board may accept from a person facing action for non-observance under the law, voluntary undertaking in respect of any matter related to the observance of any provisions of the Act at any stage of the proceedings before the Board.[91] (ii) Central Govt. has the power to call for information from the Board or Data Fiduciary and authorize blocking of access to the public which enables the Data Fiduciary to carry out activity in India.[92]

J. Jurisdiction of the Board and Penalties

The Board is subsumed with all the powers of a Civil Court and bars the jurisdiction of any other Civil Court in the country from entertaining any proceedings or granting any specific reliefs, which the Board is empowered under the Act, or any rights exercised by it thereof. The Board is also empowered to impose monetary penalties to the extent of INR 250 Crores, after the adjudication of any matter, after considering seven factors[93] enumerated under the Act.

IV. A HALF-BAKED STRUDEL

According to Auguste Escoffier, School of Culinary Arts, a successful baker is one who, *amongst* other qualities, understands the importance of "*mise en place*", communicates clearly and gives attention to details. Any attempt to cut corners, or a haphazard approach can lead to subpar results. A legal framework is no different. The efficacy of any statute largely depends upon four factors - the legislative approach which is reflected in the provisos and figuratively forms '*mise en place*' of the ensuing law; clarity and precision; flexible yet predictable and one which inspires public trust and legitimacy. Deviate from the said recipe and one is bound to end up with a law, which is no better than a soggy strudel.

Unfortunately, close scrutiny of the Act reveals the deep fault lines it hides within and the host of challenges, issues and implications, which can barely be fathomed at this moment. Following is some of the challenges and issues.

A. Flawed Legislative Approach

In the words of Chief Justice Dr. D.Y. Chandrachud, the formulation of a data privacy framework is a complex exercise which needs to be undertaken by the State

---

[90] *Section 28*.
[91] *Section 32*.
[92] *Section 37*.
[93] *Section 33(2)*.

after a careful balancing the privacy concern and legitimate state interest.[94] The Act though is not a reflection of such a thought process. While the US legislators followed the '*laissez*-faire' approach, EU followed the comprehensive data governance approach, and some Asian countries like Singapore and Japan adopted the '*disparate approach*'; Indian legislatures have adopted an '*interventionist approach'*. The fundamental flaw of such an approach lies in the fact that it not only undermines the theory of separation of power[95] and the delegation doctrine[96] but also prone to the following risks – (i) the legislature tends to delegate the "*essential legislative powers"* under the garb of delegated legislation; (ii) lacks legislative policy; and (iii) the executive branch tends to usurp the legislative powers. The DPDP Act is a classic example of this fallacy, as the Central Government has been granted wide discretionary powers without adequate legislative policy or standards under the Act, e.g. notifications of significant Data Fiduciary[97], processing of personal data by the State or its instrumentalities,[98] processing of personal data of children,[99] non-application of certain provisions under the Act to certain Data Fiduciary or class of Data Fiduciaries[100], establishment of the Board, composition and appointment of members [101], and power to call for information and issue directions.[102] By failing to set essential and clear legal policy on a whole range of issues, as mentioned above, the DPDP Act simply transforms the Executive into the primary lawmaker on multiple counts. This approach of the Central Govt. not only undermines the role played by the Legislature in the country but also casts a dark shadow on the aspiration and trust reposed on the State by billions of people in the country.

B. Allied Laws

DPDP Act is being hailed as the first cross sectoral law in the country. In reality, it is a disguised '*laissez*-faire' approach as evidenced by Section 38 of the Act which prescribes that the *Act shall be an addition to and not in derogation of any other law for the time being in force*. What it implies is that instead of having an overarching effect it encourages grandfathering of existing laws across multiple sectors. The dichotomy gains prominence considering that the DPDP Act has a direct impact on about 50[103] different laws in the country, and many of such laws govern - (i) a specific sector (e.g. Information Technology, Taxation, Health, Defense, Labor, Corporate and Financial), (ii) have independent regulators and mechanisms (e.g. Reserve Bank of India ("**RBI"**), Insurance Regulatory and Development Authority in India ("**IRDA**"), Telecom Regulatory Authority of India ("**TRAI**"), Employees Provident Fund Organization ("**EPFO**") etc.; (iii) have different objective of collecting data; (iv) have specific parameters for data collection and storage; and (v) have different adjudication and grievance mechanisms. This will obviously lead to ambiguities and inconsistencies in the way data is collected, processed, stored, safeguards and rights of the data principles and breaches adjudicated in future. While the Act attempts to amend some

---

94 *Supra Note* 58.
95 *Ram Jawaya* vs. *State of Punjab*, A.I.R. 1955 SC 549.
96 *In re The Delhi Laws Act, 1912, The…* vs. *The Part C States (Laws) Act, 1950,* A.I.R 1951 SC 332.
97 *Section 10*.
98 *Section 17(2)(a)*.
99 *Section 9(5)*.
100 *Section 17(5)*.
101 *Section 18*.
102 *Section 36 and 37*.
103 *Annexure- C, Supra note* 66.

glaring inconsistencies in about four statutes[104], but it is a long way before any uniformity or cohesiveness is achieved across sectors, if at all it is achieved or meant to achieve.

C. Applicability

The DPDP Act aptly encapsulates both territorial and extra-territorial jurisdiction to protect the data of the Data Principal. However, where it falters is the imbedded ambiguity under the Act and lack of a forward-looking approach. As discussed earlier, the Act is applicable to all individuals, who may be Indian citizens, non-resident Indians and foreign citizens, if the data is collected in India. This may pose a peculiar problem, if the individual was an EU or a New Zealand citizen, where on account of their home country laws, more than one State may exercise jurisdiction over any matter, invoke conflict of laws, and Courts will be left to determine which State has a more 'substantial connection' to the issue at hand and thereby exercise jurisdiction. Secondly, the extra-territorial application is subject to the caveat that the processing is in connection with 'activity related to the offering of goods and services to Data Principals within the territory of India'. Exercise of this prescriptive jurisdiction is predominantly based on the principle of territoriality rather than passive personality. This is a narrow application of the concept of extra-territoriality and a stark departure from most of the privacy laws across the world, e.g. GDPR, PIPL, LGPD, PDPA etc., that seeks to protect the data of individuals within their territory, irrespective of where the collection and processing is done and devoid of any such caveats, as stipulated under the Act. Thirdly, the DPDP Act adopts a traditional approach (i.e. activity related to the offering of goods and services) in its application and abjectly overlooks activities like analyzing, profiling and evaluating behavior and activities of the individuals, which is the new information gold mine. Fourthly, though the Act does not have a retrospective effect, its applicability, without any existing legislative policy or safeguards, is bound to have retrospective implication or obligations upon Data Fiduciaries across various sectors. What it implies is any processing (from mere storage and indexing to complicated analysis) of personal data, irrespective of when it might have been collected, will be within the ambit of the Act. E.g. The DPDP Act may be applicable to a bank which may be collected personal data 20 years ago from a customer for the purpose of account opening and the account is still active today. Lastly, considering the cross-sectoral applicability, the legislature has adopted a transitional approach in the DPDP Act. This approach is intended to bridge the gap between the commencement of the Act and its operation prior to it, with the objective of having a smooth transition over time from the existing laws. Unfortunately, this approach attracts a *Staling* effect. Just like a batch of freshly baked strudel loses its freshness due to retrogradation of starch molecules, inclusion of multiple transitional clauses in the Act is also likely to cause severe chaos and confusion leading to multiple litigations. This will impair effective implementation of the Act and be more disruptive than contemplated by the legislature. Consequences of similar approach being adopted in some in some the recent Indian statutes, e.g. The Insolvency and Bankruptcy Code[105] and Goods and Service Tax[106], is clear evidence of such proposition.

---

[104] *Section 44*.
[105] Act 31 of 2016.
[106] Act 12 of 2017.

D. Government—the Unregulated Hand

The DPDP Act, despite its patent objective, has been quintessentially curated as a powerful tool in the Government's armory. State and its instrumentalities have been completely kept outside the purview of the Act, in the interest of sovereignty, integrity, national security, relation with foreign states, maintenance of public order, prevention of cognizable offences[107], research, archival and statistical purpose[108]. This wide category will ensure that the Government has complete autonomy in the collection, processing, usage and storage of personal data. Currently, public entities and Government agencies hold personal information for a majority section of society. This seriously puts a dent on the transparency and trust factor, as Government excessiveness cannot be ruled out. This is a stark departure from some of the privacy laws enacted across the world e.g. GDPR, NZPA[109], PDPL[110], that encourages transparency and accountability for all Data Fiduciaries, irrespective of whether public or private sector. Secondly, the Act empowers the Government to make 'Rules"[111] for all or any aspect of the privacy laws. Unlike countries like US, UK and Australia, which have overarching legislation regulating the framing of subordinate legislation, India has none. This gives unfettered rights to the Government to alter the design of the statute itself. Thirdly, Government has been provided wide discretionary and unguided rule-making powers with respect to granting exemptions[112] under the Act, without any legislative policy or safeguards, whatsoever. This again grants wide unfettered right to the Govt. which is prone to political will, bias, and misuse. Lastly, the Govt. has complete control over the functioning of the Board, including but not limited to, the composition, appointment, terms of employment and salary of the Chairperson/Members, manner of reporting breaches to the Board and techno-legal measure to be adopted. In essence, the invisible and unregulated hand of the Govt. will control and govern the complete functioning and outcome of the Board as the Act simpliciter reduces it to a mere extended arm of the Government.

E. Data—the Modern Gold

What apple is to an *apfelstrudel*, data is to privacy laws. The DPDP Act adopts a broad-brush approach and provides for an inclusionary definition of 'personal data',[113] without any exhaustive or indicative list, of what is, will or may be considered as a personal data. This pertinent question has perhaps been left at the discretion of the Board or under the wide 'rulemaking' authority of the Government. Whatever the reasons are, the Act faulters on this aspect on multiple counts. Firstly, the definition lacks reference to '*unique identifiers*'[114], as reflected in many foreign privacy

---

[107] *Section 17(2)(a)*.
[108] *Section 17(2)(b)*.
[109] Privacy Act, New Zealand, 2020.
[110] Personal Date Protection Law, Saudi Arabia, Issued pursuant to Royal decree No. (M/19) dated 16/09/2021
[111] *Section 40*.
[112] *Section 17(3), Section 17(5), Section 9(5).*
[113] *Section 2(t)*.
[114] California Consumer Privacy Act of 2018, effective 1/1/2024 – AB 947 and AB 1194 updates posted to cppa.ca.gov April 2024. Pursuant to definition (aj) of CCPA- "Unique identifier" or "unique personal identifier" means a persistent identifier that can be used to recognize a consumer, a family, or a device that is linked to a consumer or family, over time and across different services, including, but not limited to, a device identifier; an Internet Protocol address; cookies, beacons, pixel tags, mobile ad identifiers, or similar technology; customer number, unique pseudonym, or user alias;

legislations (e.g. CCPA[115], PDPA, PIPA[116]) which constitutes the forage for marketing and solicitation activities across sectors. In absence of unique identifiers, marketing tools operating on unique IDs and cookies such as analytics platforms and customer data platforms may be successful in circumventing the law. Secondly, lack of identifiers would mean exclusion of quasi-identifiers as well, which when combined with other identifiers can render greater harm to an individual. Latanya Sweeny in her work has shown that neither gender, birthdates or postal codes uniquely identify an individual but when combined can sufficiently identify 87% of individuals in the US.[117] Thirdly and most importantly, the Act neither defines 'sensitive personal data' nor provides a segregation from aforesaid definition. 'Sensitive personal data' like biometric, financial or health data, passwords, religion etc., has been expressly defined and included under multiple privacy legislations across the world (e.g. CCPA, CPPA, PIPL, NFADP, PDPL, PDPA, PIPA etc.), owing to enhanced security requirement, active consent requirement (as envisaged under GDPR and LGP) and right to limit the usage of such data. In short, the Act neither defines sensitive data nor puts any kind of incremental obligations upon Data Fiduciaries or Processors, to safeguard the interest and privacy of the individuals the Act proclaims to protect.

F. Consent

Consent has been viewed as an expression of an individual's autonomy or control, which has the consequence of allowing another person to legally disclaim the liability for acts which has been consented to.[118] Notice coupled with choice which culminates into a consent, forms the very foundation of the consent philosophy on which the DPDP Act has been constructed. While countries across the globe are adopting a right-based approach ('Opt-in/opt-out') to privacy laws (e.g. CPPA, LGPD, PDPA) Indian legislatures are still stuck to the traditional approach. Sadly, this consent-based approach is outdated in the wake of internet, AI, and change in technology. The different elements[119] (free, specific, informed, unconditional, unambiguous with a clear affirmative action, specified purpose for processing) of consent stipulated under the Act fades away on account of the following: (i) absence of consent standards for online or digital collection, use and disclosure of personal information; (ii) absence of model notices to demonstrate compliance with the Act; (iii) discretion of the Data Fiduciary to obtain consent in any form or manner, leads to inherent weakness (e.g. pre-ticked checkboxes, notice not provided prior to processing) which are fairly common in the Indian market; (iv) lack of informed consent (e.g. non-disclosure of consequences of collection, use and disclosure of personal information or name of third parties with whom such information is shared); Secondly, this approach will invariably lead to

---

telephone numbers, or other forms of persistent or probabilistic identifiers that can be used to identify a particular consumer or device that is linked to a consumer or family. For purposes of this subdivision, "family" means a custodial parent or guardian and any children under 18 years of age over which the parent or guardian has custody.

115 *Supra Note 127.*

116 Personal Information Protection Act, South Korea, Act No. 19234, March 14, 2023.

117 L. Sweeney, *"Simple Demographics Often Identify People Uniquely",* Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

118 Adam Moore, Toward Informational Privacy Rights, 44 San Diego Law Review (2007) at p. 812; Anita L. Allen, Why privacy isn't everything: Feminist reflections on personal accountability (Rowman & Littlefield, 2003) at pp. 115-16; John Kleinig, The Nature of Consent in The Ethics of Consent- Theory and Practice (Alan Wertheimer and Franklin Miller (eds.), Oxford University Press, 2009) at p. 4.

119 *Section 6(1)*.

consent fatigue (apart from dampening user experience) for the Data Principal, in absence of a consolidated and supervised consent dashboards or platforms, which are seamlessly integrated with consent manager or tag managers. Thirdly, one of the crucial elements overlooked (intentionally or inadvertently) in the Act is the relationship between consent and contractual necessity. Both the ingredients needed to be de-coupled, as any consent to processing extracted by holding contractual rights hostage tantamount to consent being treated as 'not' free.[120]. Surprisingly, the Act does not include contractual necessity under 'Legitimate uses'[121] and it has been left to the Government, Board or Courts to clarify whether businesses can enforce contract under the aforesaid Section or will it again be required to obtain explicit consent for processing of personal data.

G. Non-Consensual Processing

Any free and fair normative privacy framework is dependent on two factors, i.e., autonomy of individual vis-à-vis national, social, and economic interest. The DPDP Act which rechristened the concept of 'deemed consent' (introduced under the Bill of 2022), as "Certain legitimate Uses"[122] has completely dislodged this fine balance. Firstly, the Act places the State and its instrumentalities on a lower pedestal as compared to private entities, especially in relation to processing data on a non-consensual basis. It includes scenario, where consent has *either* been previously provided[123] or no-consent has been provided at all[124]. This wide exception is bound to raise several interpretational issues e.g. nature of entity performing the function, nature of the function itself and the extent of usage of such data. Secondly, the legitimacy of collection and usage of such data will be opaque, considering that the State and its instrumentalities are empowered to formulate the policy under delegated legislation, in absence of any specific provision in the Act. Thirdly, no thought has been given to data minimization, purpose limitation and transparency, as suggested by the Supreme Court of India in the *Puttaswamy Judgment*. On the contrary, State excessiveness will be prone to occur in lieu of the consideration (subsidy, benefits, services, permits etc.) promised to the Data Principal. Lastly, the so called 'legitimate uses' lacks any sort of safeguards to protect the interest of the Data Principals. E.g. usage of data for employment purposes should be invoked only when it involves a disproportionate or unreasonable effort on the part of the employer to obtain a valid consent.

H. Data Fiduciary

Data Fiduciary plays a pivotal role in any privacy law considering the dual objective it seeks to fulfil, i.e., collection, usage, and storage of data *vis-à-vis* the accomplishment of the purpose desired by the Data Principal. The DPDP Act, while encompassing most of the obligations envisaged under the 2019 Bill, lacks the comprehensiveness and finesse, to shape India's digital landscape in the 21st century. One of the foremost issues is the absence of obligation to ensure fair and reasonable processing by the Data Fiduciary to prevent abuse of power. Fair[125] and reasonable processing implies two elements, i.e., obligation to uphold the best interest of the Data

---

120 Recital 3, GDPR.
121 *Section 7*.
122 *Supra Note 134.*
123 *Section 7(b)(i)*.
124 *Section 7(b)(ii)*.
125 UK GDPR, Article 5(1)(a); Recital 39 GDPR.

Principal and the processing is not beyond the expectation of the Data Principal.[126] Secondly, the Act lacks explicit provisions for data minimization when compared to privacy laws[127] across the globe. The data limitation principle forms the bedrock of any privacy laws[128] as it ensures that the data collected is limited to what is necessary to achieve the primary purpose and if such collected data is no longer necessary for such purpose, it ought to be destroyed. Unfortunately, in absence of any explicit provision under the Act, the Data Fiduciary is under no such express obligation and thereby increasing the chances of abuse of power. This risk attains a larger magnitude and leads to tangible harm to individuals with the emergence of Big Data, processing vast amount of data at scale to discern patterns of individual patterns and market trend.[129] Thirdly, the Act does stress on the transparency principle. Mere obligation on the Data Fiduciary to give notice to the Data Principal at the time of collection is highly inadequate. Additionally, the obligation to inform the Data Principal of the basis of processing, legal obligation for such processing, persons with whom the data is shared, and period of retention is entirely absent. Lastly, the Act does not deal with the principle of storage limitation, which ought to have obligated the Data Fiduciary to delete or anonymize the personal data after the purpose is achieved. This essentially exposes the personal data to theft, copying, transferring or usage, without any kind restriction or consequences upon the Data Fiduciary.

I. Missing Ingredients of the Act

The DPDP Act has, consciously or otherwise, watered down various provisions which were incorporated in the Bill of 2019. Some prominent provisions need a special mention here. The Act entirely excludes offline personal data and data collected through non-automated processing from within its purview. Data portability[130], which enables the Data Principal to receive structured format of the collected data and transfer it to a different institution has been completely discarded. While the Bill of 2019 proposed deanonymization as a criminal offence, the Act has completely de-criminalized all offences. Concept like 'Right to be forgotten" has been outrightly junked considering competing State rights and interests. Complete autonomy has been bestowed upon the Government (without any requirement to consult or seek guidance from the Board) in deciding and notifying the countries where personal data may be transferred for the purpose of processing. The 'harm' caused on account of processing the data, including the obligations on the Data Fiduciary to mitigate such harm or the right of the Data Principal to seek compensation for such harm are entirely missing in the Act. Lastly, the Act encompasses various generic words without suitable explanations or standards, leaving it for the Government, Board, or the Courts to define, interpret and implement the same. E.g. "detrimental effect",[131] "well-being of the child"[132], or "verifiably safe".[133]

---

126 *Supra Note 66.*
127 UK GDPR, Article 5(1)(c); Recital 39 GDPR, PDPL.
128 GDPR, FIPPs, OECD Guidelines on the Protection of Privacy and Transborder flow of Personal Data (2013).
129 White Paper of the Committee of Experts on a Data Protection Framework for India, 2018, at p. 8.
130 Incorporated in various privacy laws. E.g. CDPA, LGPD, PIPA.
131 *Section 9(2)*.
132 *Ibid*.
133 *Section 9(5)*.

J. A Toothless Board

What a good baking instrument is to a strudel, a robust enforcement mechanism is to a good law. Compromise on the instrument and the result will be a streusel and not a strudel. The DPDP Act suffers from this major lacuna, as the Board (in comparison to DPA-its proposed predecessors) has been reduced from a sector-agnostic, independent and comprehensive regulatory body to a procedural body merely for the purpose of overseeing data breaches, direct remedial measures and conduct inquires. Firstly, the independence of the Board has been seriously compromised given that the selection and appointment of the Chairpersons/members, their tenure, salary, allowances, and functioning are exclusively decided by the Central Government[134] and not by a specialized committee. No transparent pre-requisite regarding the professional qualification, expertise or experiences have been prescribed. Secondly, the possibility of conflict-of-interest situations, arising vis-à-vis the Chairpersons/members and their functioning cannot be ruled out in absence of a clear demarcation on the powers and functionality coupled with the complete discretion available with the Chairperson under the aegis of the Central Government. Thirdly, the Board does not have any regulatory tools like formulation of best practices code, issuance of guidance or public statements. All is at the behest of the Central Government. Lastly, the Board ought to have been the independent regulator exercising powers across the sectors in the Indian economy. Unfortunately, the Board has a subordinate status to the various sectoral regulators, which will seriously undermine the entire privacy framework and enforcement mechanism. To sum up, the Board is a toothless tiger under the tutelage of the Government, which will be incapable of protecting the interest of the Data Principal, both in letter and in spirit.

K. Dysfunctional Dispute Resolution Mechanism

One of the gaping drawbacks under the Act, is empowering TDSAT[135], an existing tribunal to discharge the functions of the appellate tribunal under the Act. TDSAT, a twenty-four-year-old tribunal, is already the appellate tribunal for telecommunications, cable services, broadcasting, cyber and airports related disputes. It is saddled with 5426 pending cases[136] (approx. 50% of the total disposed cases[137] since inception) as on date. Hence, it is highly improbable that TDSAT will render effective and speedy adjudication mechanism. Secondly, the TDSAT comprises of members who have no or little technical expertise, know-how or experience in the field of information technology, cyber, internet laws, AI or such related fields. Hence, to think that TDSAT will have any meaningful impact or adjudication role under the Act is extremely farfetched by any yardstick.

L. Technologically Agnostic

Technology and privacy, quantitively speaking, are inversely proportional to each other, in terms of growth, implication, risk and protection. Hence, privacy laws of

---

[134] *Section 19, 20, 22, 27(3)*.

[135] Telecom Disputes Settlement and Appellate Tribunal established under the Telecom Regulatory Authority of India Act, 1997.

[136] Statistical Report of Pending Cases (2024), available at: https://tdsat.gov.in/Delhi/services/pending_report.php.

[137] Total Cases-11280. Statistical Report of Disposed Cases (2024), available at: https://tdsat.gov.in/Delhi/services/disposal_report.php.

the 21st century need to have a harmonic relationship with technology. Unfortunately, the DPDP Act does nothing in this context. Firstly, biometrics and genetic data, which includes fingerprints, retina, voice and facial patterns and genetic code have not been provided enhanced security, safety and storage safeguards unlike privacy laws of many countries[138]. Secondly, surveillance mechanisms engaged by the State and private entities, e.g. CCTVs, and GPS, and enhanced technological features like night camera, motion detection and computer assisted operations, are completely outside the realm of the Act. Thirdly, the Act does little to address the issues arising out of Data mining and Internet, especially considering the tremendous amount of data which is collected, stored and processed without the consent and knowledge of the users. E.g. Cookies coupled with spam and spyware facilitate the collection and analysis of personal information leading to identification of individuals, without the user even realizing or knowing the consequences of it. Lastly, the Act is extremely ill equipped to handle some of the future technologies like ambient technology, neurolinguistics and grid technology, as these complex technologies are capable of not only collection and monitoring of personal information of the user but also effecting changes in behavioral and neural patterns, choices and responses of such users.

V. TOWARDS A ROBUST AND SUSTAINABLE FRAMEWORK

Considering that the DPDP Act has been enacted after nearly 81 amendments since it was first tabled before Parliament in 2019, a complete overhaul now may not be possible or advisable at this time. Hence, it is felt that to achieve a robust and sustainable privacy framework within the ambit of the Act, the following 14-point recommendations needs to be considered to meet the aspirations of digital India and its 1.5 billion inhabitants:

1) The Central Government should extend the application of Act to cover processing of all data pertaining to individuals within the Indian territory, irrespective of where the collection and processing is done or whether the Data Processor is established or whether such Data Processor is providing goods and services in India. Additionally, all analyzing, profiling and evaluating behavior and activities of the individuals, should be brought within the ambit of data processing.

2) To minimize the effects of transitional approach of the Act, the Central Government needs to provide comprehensive Rules and specific timelines for its implementation, in consultation with sectoral regulators and stakeholders.

3) To minimize the effect of consent fatigues for the Data Principles, the Central Government must provide or facilitate Consent frameworks, consent dashboard and data trust score for Data Fiduciaries, operated by public or regulated entities.

4) The Central Government should formulate prescriptive rules related to consent from Data Principals, especially in connection with consent

[138] GDPR, CPRA, nFADP, PDPL.

managers and children's data, by stipulating models forms and dynamic consent renewals and withdrawals.

5) The Central Government should formulate rules and guidelines on collecting, usage and storage of sensitive data and corresponding enhanced obligations upon data fiduciaries. Opt-in rights and express consent must be made mandatory.

6) The Central Government should formulate adequate Rules to protect "Big Data" processing by Data Processors based on collection and purpose principles. Perhaps usage of blockchain technology can be considered in this context.

7) The Central Government along with the sectoral regulators should formulate rules and guidelines on data portability and interoperability, especially in sectors facing increasing digitization e.g. banking, insurance and social media. Introduction of hybrid and multi-cloud strategies coupled with uniform terminology and standards for data portability and interoperability, will go a long way to mitigate the risks arising from such activities.

8) The Central Government should formulate rules and guidelines related to the usage of surveillance cameras, workplace surveillance, outsourcing services and direct mailing by Data Fiduciaries or Data Processors.

9) To avoid multiple litigations and disputes related to data privacy issues and breaches by the State and its instrumentalities, the Central Government must formulate and implement a grievance mechanism for Data Principles in the country.

10) The Central Government in consultation with sectoral regulators/experts should frame and promulgate rules, code of best practices, policies and advisory related to data discovery, data mapping, loss prevention, data erasure and recovery.

11) The Central Government and Board should encourage, and facilitate usage of privacy enhancing technologies (PETs) e.g. secure multiparty computation, differential privacy computing and on-device analytics, in addition to the existing technologies like encryption, and tokenization, across sectors.

12) Specialized bench should be constituted under TDSAT, consisting of tecno-legal experts from various sectors. The functioning of the Bench should be on a 'fast-track' basis and equipped with modern technology to ensure live streaming of proceedings, digital filings, recordings and proceedings, 24x7x365 accessibility. Additionally, powers may include appointment of amicus curia and technical experts, pre-mediation process, disclosure requirements and fixed timelines for expediating the disposal of cases.

13) The Board should enter into formal arrangements or MoU with sectoral regulators to lay out a uniform and comprehensive plan or framework for the implementation of the Act, including addressing critical elements of the privacy laws, e.g. form and manner in which personal information is collected, stored and used, breaches and grievance mechanism.

14) Sector regulators and industry bodies, e.g. Indian Bank's Association (IBA) for banking, should formulate comprehensive plan or framework for the implementation of the Act, to avoid information and implementation asymmetry within the respective sectors.

CONCLUSION

Given the deep fault lines in the Act, as demonstrated above, it will be interesting to see how and when the Act is implemented, Rules promogulated and its comprehensiveness to design the privacy framework for digital India. The regulatory architecture and the institutional framework that will crystallize over the next few years will decide how well (or not) personal data privacy is safeguarded. The new law provides for the dough but is clearly far away from a *de jure* data privacy law. Lastly, considering the '*interventionist approach'* adopted by the Government, the success of the Act will vastly depend upon the political will, intention and proactiveness of the government machinery. Until then the DPDP Act is simply a strudel served raw!

A VISION FOR DIGITIZING JUDICIAL PROCESSES AND INTEGRATING ARTIFICIAL INTELLIGENCE IN PAKISTAN'S JUDICIARY: ENHANCING EFFICIENCY AND UPHOLDING JUDICIAL INTEGRITY

Faiza Khalil*

Abstract: The judiciary, as the cornerstone of justice and the rule of law, is at a pivotal juncture to harness the transformative power of digital technology and artificial intelligence (AI) to enhance its operations. This essay outlines a comprehensive approach for digitizing judicial processes in Pakistan, incorporating AI integration by drawing parallels with successful international model. The focus is on the need for systemic change to ensure efficiency, transparency, and accessibility in the legal system. Current challenges include a lack of proper implementation of the rule of law, prolonged trials, and low public confidence. Traditional methods of case filing are manual and paper-based, leading to inefficiencies. The essay proposes a step-by-step transformation starting from e-filing to the digitization of the entire lifecycle of a case, aiming to modernize Pakistan's judiciary. AI can aid in legal research, evidence standards, and sentencing, offering predictive capabilities and streamlining routine case management. Ethical considerations and the need for human judicial discretion are emphasized to balance AI's assistance with maintaining judicial integrity and fairness. This digital transformation can restore public trust and efficiency in Pakistan's judiciary, paving the way for a modern, digital legal system.

Keywords: Technological Integration in Courts; Predictive Analytics in Law; Transparency in Judicial Processes; Judicial Modernization; Artificial Intelligence in Judiciary; Digital transformation; AI and Legal Ethics; Case Management Systems; E-Filing; Legal Tech Innovation; AI in Legal Research; Efficiency in Judiciary; AI and Evidence Standards; AI in Sentencing; Legal Information Systems

* University of Karachi, Pakistan.

Table of Contents

Introduction 110

I. Current Challenges in Pakistan's Judiciary 110

II. Traditional Case Filing Methods 110

III. The Need for Digital Transformation 111

IV. Integration of AI in Judicial Processes 112

V. AI in Legal Research and Decision-Making 112

VI. AI in Sentencing and Evidence Standards 112

VII. AI in Routine Case Management 113

VIII. The Chinese Model of Smart Courts 115

A. Historical Development 115

B. Initiatives For the Transformation 116

C. Procedure of the Courts 117

IX. Ethical Standards and Transparency 117

X. Practical Considerations for Judicial Use of AI 118

XI. Strategic Roadmap for Transforming Pakistan's Judicial System Inspired by the Chinese Model of Smart Courts 119

Phase 1: Foundational Assessment and Strategic Planning 119

Phase 2: Technology Integration in Judicial Processes 119

Phase 3: Enhancing Judicial Transparency and Efficiency 120

Phase 4: Scaling and Integrating Advanced Technologies 120

Phase 5: Continuous Improvement and Innovation 121

Conclusion 121

INTRODUCTION

No society may be esteemed an ideal civilization, unless it venerates and abides by the laws. The rule of law may only be established when the populace holds unwavering faith in the sanctity and rectitude of law. Moreover, the public must be so instructed that they comport themselves according to the law without coercion. Whilst all pillars of the state hold significant import, the judiciary is preeminent in its charge of dispensing justice. Justice, in its most unadorned form, signifies the rightful placement of all things. The judicial institution serves as the vigilant sentinel, ensuring none are met with injustice.

I. CURRENT CHALLENGES IN PAKISTAN'S JUDICIARY

In our realm, due to the lack of proper implementation of the rule of law and the protracted nature of trials, litigants have lost faith in the judicial apparatus, and are often compelled to settle disputes or chastise wrongdoers outside the bounds of court, rather than endure interminable waits for adjudication. Rectifying these deficiencies will augment the populace's access to and confidence in the courts, thus disrupting the cycle of mistrust and apathy towards the judiciary, and aiding the judiciary in reclaiming its esteem.

The judiciary faces significant challenges as reflected in the 2023 World Justice Project (WJP) Rule of Law Index, where it ranks 130th out of 142 countries. The country shows weak checks and balances on the executive branch by the judiciary and legislature, leading to significant executive overreach and limited accountability. Corruption is pervasive, affecting the judiciary, police, and public services, with ineffective anti-corruption measures in place. Transparency and accessibility of government operations are limited, with insufficient publicized laws and poor civic participation. Fundamental rights, including freedoms of speech, assembly, and association, are inadequately protected, and human rights violations are common. Despite some improvements in security, issues such as terrorism, violence, and crime persist, impacting the overall law and order situation. The civil justice system is plagued by delays, inefficiency, high case backlogs, and procedural complexities, which undermine public trust. The criminal justice system also faces significant issues, including inadequate protection of due process, prolonged pre-trial detentions, and poor prison conditions, leading to low public confidence.[1] The number of judges serving is around 4200 including superior judiciary to handle the backlog of 2.1 million cases.[2]

II. TRADITIONAL CASE FILING METHODS

Traditional or obsolete methods of case filing, which are largely manual and paper-based, still exist in Pakistan. Lawyers or individuals involved in the case prepare necessary legal documents manually. Once the documents are prepared, the litigant often needs to be verified through the biometric system of NADRA. The next step involves physically submitting these documents to the court's registry or filing office. The person filing the case would need to go to the court, often waiting in long lines for

[1] World Justice Project, WJP Rule of Law Index 2023, https://worldjusticeproject.org/our-work/research-and-data/wjp-rule-law-index-2023 (last visited June 16, 2024).
[2] Law & Justice Commission of Pakistan and National Judicial Policy Making Committee, Judicial Statistics of Pakistan 2022 (2022).

their turn to submit their documents to the filing clerk. After the case is filed, notifications are sent out manually to the parties involved in the case, often via postal mail, informing them of the case details, court dates, etc. While online notices have begun to be issued, they are not being totally disseminated through official accounts or numbers. Accessing case files or past records require physically going to the court and requesting to view the files, which are stored in large filing cabinets or rooms dedicated to archives. The current process required for a lawyer to obtain a certified copy of a case file or order involves a tedious and inefficient system where documents are manually searched, transported, and copied, involving multiple departments and requiring physical presence and payments at various stages. This manual method is time-consuming, labor-intensive, and susceptible to issues like loss of documents, delays, and lack of transparency.

III. THE NEED FOR DIGITAL TRANSFORMATION

Modern judicial systems are increasingly moving towards not only electronic filing (e-filing) systems but to robotic courts which streamline these processes, reduce paper use, improve efficiency and accessibility as well as retain the capability to handle the case backlog. [3] The advanced issues with respect to robotic courts like unpredictability, complexity and legal personhood are under discussion. Emergent behavior in AI systems, where they exhibit actions not explicitly programmed due to learning from vast amounts of data, can lead to unforeseen outcomes. It is debated that such unpredictability makes it difficult to determine responsibility when AI causes harm or violates laws, as traditional liability frameworks based on foreseeability are mostly inadequate. Alongside AI systems are often collaboratively developed by multiple entities, complicating the assignment of causality and responsibility for failures or harmful actions. There is another ongoing debate on whether AI should be granted legal personhood to simplify liability issues, making the AI itself responsible. Proponents suggest new standards for foreseeability and accountability, and mechanisms to hold multiple parties jointly responsible when their contributions lead to harm.[4]

Reverting to the aspect of modernization of judiciary in the country, the necessary technology and human resources for the digitization and AI integration, to some extent are already available. The tasks outlined can be achieved in a relatively short timeframe, paving the way for a more profound transformation towards a modern, digital judiciary. This journey must begin with a step-by-step transformation of each judicial process, starting from e-filing to the digitization of the entire lifecycle of a case. Allowing digitally uploading petitions along with Computerized National Identity Cards (CNICs), biometric data, and photo identity are similar to the process already in place for online passport renewals for the Pakistani diaspora. All court documents and records shall then be integrated into a central database. Lawyers or litigants could access and download needed documents through a secure online system after

[3] Celine Cousineau, Smart Courts and the Push for Technological Innovation in China's Judicial System, CSIS (Apr. 15, 2021), https://www.csis.org/blogs/new-perspectives-asia/smart-courts-and-push-technological-innovationchinas-judicial-system. (Last visited June 05, 2024).

[4] P.W. Grimm, M.R. Grossman, S. Gless, & M. Hildebrandt, Artificial Justice: The Quandary of AI in the Courtroom, Judicature Int'l (Sept. 2022).

undergoing necessary checks and electronic payment, thereby reducing the turnaround time from days to minutes.[5]

IV. INTEGRATION OF AI IN JUDICIAL PROCESSES

This digital transformation will also lay the groundwork for integrating artificial intelligence (AI) into the field. Currently, the technology's potential to benefit the judicial system is untapped due to the absence of such modernization. The technology could facilitate by sifting through an immense corpus of legal documents, including statutes, case law, and secondary sources, at a pace and with a level of accuracy that far exceeds human capabilities. It could also be used for summarizing and comparing texts. AI-driven analysis can flag inconsistencies or fallacies in argumentation, contradictions in witness testimonies or flaws in the application of legal principles.[6]

V. AI IN LEGAL RESEARCH AND DECISION-MAKING

By utilizing natural language processing, AI can interpret the query of a judge and return pertinent information, drawing from a wide range of jurisdictions and legal systems. This ensures that judicial decisions are based on the most recent and relevant legal precedents. For instance, AI-powered legal research tools can analyze the fact patterns of a case and identify analogous cases across numerous jurisdictions, as well as highlight seminal cases that have significantly influenced the interpretation of specific laws. This gives judges comprehensive insights into the legislative framework and judicial reasoning. AI systems can also track changes in legal trends, societal norms, and judicial interpretations over time. By identifying shifts in how courts are interpreting laws and applying precedents, AI can help signal when a precedent may need to be revisited or reconsidered, thus maintaining the balance between stability and evolution.

The predictive capabilities of AI can yield tentative forecasts about the outcomes of litigation based on historical data and mining patterns from a vast array of similar cases. While no algorithm can guarantee the result of a legal case due to the myriad variables and the intrinsic unpredictability of law, AI can nonetheless unearth statistical trends that may offer judges substantive insights. Judges can then employ these predictions to weigh the purported strengths of arguments, to inform their understanding of how similar cases have been adjudicated, and to gain a preliminary sense of the likely direction a current case may take.

However, this functionality requires access to legal decisions that are often behind firewalls and not publicly available. There needs to be a shift in how legal judgments are accessed, removing barriers to make these important resources part of the machine learning dataset. This step is crucial for the advancement and efficacy of AI tools like Judge GPT in the judicial system.

VI. AI IN SENTENCING AND EVIDENCE STANDARDS

Formulating evidence standards with the help of AI could be particularly beneficial in adjudicating cases. Leveraging AI to adhere to these standards can

---

[5] . Naeem Sadiq & Zulfiqar Ali Qureshi, The Express Tribune, Feb. 15, 2023.

[6] . J. Shafiq, H.M.S. Shafiq, & M.S. Sarwar, Use of ICTs and Artificial Intelligence to Overcome Judicial Trial Delays in Pakistani Courts, 6 Pakistan Lang. & Humanities Rev. 1153 (2022).

significantly enhance the verification and comparison processes. AI can systematically ensure that evidence meets the required standards by excluding defective evidence, which may be unreliable or irrelevant, and illegal evidence, which could compromise the integrity of legal proceedings. By automating these checks, AI minimizes the risk of human error and bias. Additionally, AI can help safeguard the process from external interference, ensuring that the evidence remains uncontaminated and that the evaluation process remains fair and objective. This technological approach not only streamlines the evidentiary process but also upholds the principles of justice and accuracy in legal and investigative contexts.[7]

Sentencing is among the most challenging tasks a judge faces, requiring a balance between legal guidelines, the severity of the offense, and mitigating or aggravating factors. The strictness of sentencing can also differ according to a judge's choices, which can in turn depend on their nature, social values and experiences. Some judges prefer minimum punishment while others deem the maximum term as appropriate. AI can aid in navigating this complexity by churning through data to deliver evidence-based recommendations on sentencing. By accounting for an array of considerations, such as the nature of the crime, prior convictions, the accused's background, and vast sentencing records, AI algorithms can help ensure consistent and equitable sentencing practices. Determining the likelihood of recidivism is a key component in decisions related to bail, parole, and sentencing, thus AI can play a crucial role by evaluating myriad variables to assess the risk a particular individual might pose if afforded certain freedoms. Utilizing data points from past behavior, social and demographic predictors, and comparison with historically similar individuals, AI provides an evidence-based framework for these assessments.[8]

VII. AI IN ROUTINE CASE MANAGEMENT

Yet there is another aspect to be benefitted from the technology. Many judgments such as default rulings and declarations of inadmissibility are generated in routine manner in a court. Numerous cases require a straightforward evaluation without a formal hearing, and others are resolved through settlements. It is only a minor fraction of the cases that present complex and contentious issues. It is essential to acknowledge that the judicial process, and consequently the requirement for information technology, is not uniform across all cases. A considerable number of routine cases feature predictable outcomes. In such instances, the court's ruling could be a document produced predominantly through an automated process, based on the data submitted. Here, courts could typically receive digital submissions, whereby the presenting party supplies data electronically, thus obviating the need for manual re-entry. The processing of these cases could be partially or even fully automated using artificial intelligence, precisely because the outcome is largely or entirely determinable in advance. Thus the human input could be avoided and a robotic procedure could curb the burden from the institution.[9] In our system, orders in family cases, rent disputes, bail applications, succession applications and several others could be generated from

---

[7] Gulimila Aini, A Summary of the Research on the Judicial Application of Artificial Intelligence, 9 Chinese Stud. 14 (2020), https://doi.org/10.4236/chnstd.2020.91002.
U.S. Tahura & N. Selvadurai, The Use of Artificial Intelligence in Judicial Decision-Making: The Example of China, Int'l J.L., Ethics & Tech. (2022), https://doi.org/10.55574/PYEB5374.

[8] Gulimila Aini, *Supra note 7*.

[9] A.D. (Dory) Reiling, Courts and Artificial Intelligence, 11 Int'l J. for Ct. Admin. 8 (2020), https://doi.org/10.36745/ijca.343.

AI automated setup. These Chatboats could be useful even in alternate dispute resolution.

Yet there is another aspect to ponder, the nature of the "mental process" involved in artificial intelligence systems differs markedly from the cognitive process of human beings. A mental process is a prerequisite for a legally effective decision, which a computer system inherently lacks. While AI systems do undertake a form of mental process, characterized by a "neural network," this process involves two distinct phases: first, a learning phase where data sets are gathered and trained, and second, an application phase where the system applies what it has learned. AI systems exhibit "intelligent behavior" by analyzing their environment and acting with some degree of autonomy to achieve specific goals. Despite these capabilities, while applying the technology in judicial process, it is to be kept in mind that the judicial functions require human intelligence, the ability to interact with compassion, emotion, and agile responsiveness—qualities that computer programs, to date, have not replicated. Thus, while AI can assist in certain tasks, certainly not an ultimate decision maker, the nuanced and empathetic decision-making essential to judicial roles remains uniquely human.[10]

AI decision-making systems render decisions by seeking similarities in case facts, whereas human judges consider each case on an independent basis. In AI automated setup independence of a judge could be compromised by the combined intentions of programmers, software engineers, information technology companies, and other entities.[11] The debate also persists on whether AI can be a legal personality bearing rights and obligations. The mental process of human decision-makers and AI systems also differs in the scope of the material considered and the relevant temporal parameters. In human decision-making, judges have access only to the client's legal data. In contrast, AI decision-making systems have access to all data entered by programmers and analysts, in addition to what is available to the judge. Moreover, while human judges consider both past and future events, AI judging largely depends on past events as embodied in the data sets used to train the AI system. In some cases, a formulated algorithm based on past events may not suffice to address the matters before the decision-maker. AI searches data to identify patterns to predict outcomes. Unlike AI, a human judge can be persuaded through reasoned legal argument. Arguably, AI cannot mimic general human cognition and intelligence, while humans often understand intents, emotions, and implied assumptions.[12]

Thus, AI can be a useful adjunct to the human decision-making process through the analysis of big data. It can be designed to handle simple matters independently. However, in complex matters demanding social values and choice, it is prudent to employ AI as an assisting tool to human decision-makers, rather than as independent arbiters.

In order to serve these purposes, it is imperative that knowledge map for the premise of AI is constructed in a manner which demands high standards in terms of data quality, model sophistication, and the granularity of knowledge, even for relatively

---

[10] Gulimila Aini, *Supra note 7*.
[11] M.M. Rahman, Should I Be Scared of Artificial Intelligence? Academia Letters, Art. 2536 (2021), https://doi.org/10.20935/AL2536. (Last visited May 30, 2024).
[12] Gulimila Aini, *Supra note 7.*

straightforward cases. The required level of detail in a knowledge map is considerable. The construction process involves case handlers meticulously labeling each element, a task that is both time-consuming and labor-intensive.[13] Furthermore, the complexity of legal language, adhering to the principle of stare decisis poses additional challenges. Creating a highly accurate knowledge map that faithfully captures the nuances of the case and the judge's rationale is, therefore, a daunting task.

From a logistical standpoint, AI systems adeptly handle the complexities of case management, streamlining the judicial workflow. Such systems can track the progression of cases, notify judges of upcoming deadlines, schedule proceedings, and even offer recommendations for suitable allocation of cases to different judges based on expertise and current workload.

VIII. THE CHINESE MODEL OF SMART COURTS

China being global leader, has made swift advancement in integrating legal technology into its judicial sector, driven by a higher trust in AI in East Asia compared to Western countries, and the need to address disparity between the growing number of cases and deficient workforce. This has led to the development of "smart courts" using advanced technologies such as AI, cloud computing, and big data.

A. Historical Development

Concerns about vulnerable populations and equitable access to justice prompted a focus on technological advancements. Thus the "intelligent" modernization of courts through information technology was intended to make judicial functions more transparent, efficient, and centered on people. Given the growing number of internet users in the country, technological convenience was believed to enhance judicial processes and substantially lower the costs related to accessing justice. The transformation of China's court system into 'smart' courts began in the 1990s. The significant focus was on using technology to address judicial administration challenges, moving from handwritten to word-processed documents, and setting the stage for future online filing systems. In 1997, the SPC emphasized that traditional handwritten court files could no longer meet the needs of increasing case numbers, necessitating the shift to computerized systems. This reform phase centered on digitizing court files and case management documents. According to the first Five-Year Reform Outline of People's Courts (1999-2003), all courts were expected to digitize their files by the end of 2001 and complete a nationwide internet network by 2003 to enhance judicial administration.[14]

In 2014, Chief Justice Qiang Zhou urged all Chinese courts to leverage technology to enhance the justice system, promoting public experiences of 'fairness' and 'justice' through judicial transparency. This included measures like uploading judgments online and livestreaming court hearings.[15]

---

[13] Gulimila Aini, *Supra note 7.*

[14] Qi Zhou, Guangyin Yanzhong Fayuan de Bianqian (光阴眼中法院的变迁) [The Changes of Courts in the Eyes of Time], Renmin Fayuanbao (人民法院报) [People's Court Daily], Sept. 1, 2018, 4:51 PM, https://www.chinacourt.org/article/detail/2018/09/id/3482084.shtml. (Last visited June 26, 2024)

[15] Changqing Shi, Tania Sourdin & Bin Li, The Smart Court – A New Pathway to Justice in China?, 12 Int'l J. for Ct. Admin. 4 (2021), https://doi.org/10.36745/ijca.367.

B. Initiatives for the Transformation

The phase of the transformation of China's courts, from 2004 to 2013, was characterised by the introduction of internet-assisted court hearings. Computing and internet technologies were utilised for case management and hearings during this period. Along with the emergence of internet-assisted trials in the early 2000s, hearing activities began to be recorded using audio and video technology. This stage of transformation was prompted by the Supreme People's Court (SPC) in response to justice reform specified task for courts to accomplish in the second Five-Year Reform Outline of People's Courts (2004–2008).[16] During this period, judicial transparency was significantly enhanced by livestreaming court hearings to the public. In September 2009, the Beijing High People's Court announced the launch of a city-wide livestreaming website, allowing the general public throughout the country to simultaneously watch hearings in any Beijing court. Similarly, in March 2010, the High People's Court of Henan Province in Central China reported conducting its first livestreamed hearing reflecting its commitment to improve the openness and transparency of the justice system.[17] Regulating live stream hearing was a significant task in the third Five-Year Reform Outline of People's Courts (2009-2013) also.[18]

In 2017, one year after the 'smart court' concept was officially introduced, President Xi Jinping called for the integration of modern science and technology into judicial reforms. This initiative aimed to support the development of socialism with Chinese characteristics. By 2017, it was recognised by China's executive leadership that modernising the courts required the infusion of advanced technology. The most recent Five-Year Reform Outline of People's Courts (2019-2023) confirmed that one of the ten primary objectives of justice reform is to comprehensively advance the construction of smart courts. The Supreme People's Court (SPC) has identified specific measures to achieve this goal, including the use of AI technologies, enhancements in voice-to-text applications during hearings, and the implementation of intelligent auxiliary case management systems.[19] As part of China's strategy to digitize judicial services, the Hangzhou Internet Court was established in August 2017 in Zhejiang Province, home to Alibaba's headquarters. This court handles internet-related cases, such as online shopping disputes, through its online platform. The platform allows for complete digital

---

[16] Y. Chen, The Supreme Court Issued the 25th Five-Year Reform Outline, People's Court News, Oct. 28, 2005, http://www.china-judge.com/ReadNews.asp?NewsID=3280&BigClassName=%CB%BE%B7%A8%B8%C4%B8%EF&BigClassID=17&SmallClassID=25&SmallClassName=%CB%BE%B7%A8%B8%C4%B8%EF&SpecialID=0. (Last visited June 27, 2024)

[17] Sina News, Beijing High People's Court Now Livestreaming Court Hearings, Sina News (Sept. 17, 2009), http://news.sina.com.cn/c/2009-09-17/065816311211s.shtml. (Last visited June 27, 2024) Central Government Portal, Henan Conducts the First Court Trial Webcast to Further Promote Judicial Justice, Central Gov't Portal (Mar. 11, 2010), http://www.gov.cn/gzdt/2010-03/11/content_1553005.htm. (Last visited June 27, 2024)

[18] China News, An Analysis of 10 Key Words in the People's Court's Three-Five Year Reform Outline, China News (Mar. 26, 2009), http://www.chinanews.com/gn/news/2009/03-26/1618736.shtml. (Last visited June 27, 2024)

[19] Xinhua News Agency, Unswervingly Advance Judicial Reform and Take the Road of Socialist Rule of Law with Chinese Characteristics-General Secretary Xi Jinping's Important Instructions on Judicial System Reform Aroused Enthusiastic Responses, Xinhua News Agency (July 11, 2017), http://www.xinhuanet.com//politics/2017-07/11/c_1121302631.htm. (Last visited June 27, 2024) The Paper, The Full Text of the Supreme Court's "Fifth Five-Year Reform Outline" | Authoritative Interpretation, The Paper (Feb. 27, 2019), https://www.thepaper.cn/newsDetail_forward_3051310. (Last visited June 27, 2024)

judicial proceedings, from case filing and document serving to evidence exchange, online hearings, and judgment delivery.[20]

C. Procedure of the Courts

Key smart courts, as referred above utilize AI to streamline case management, reduce delays, and cut costs. These courts are interconnected through a national e-evidence platform based on blockchain, which supports evidence authentication and examination. The AI-driven processes begin with electronic filing, where pleadings and relevant materials are scanned to produce electronic files. During trials, AI tools facilitate examination and cross-examination, with synchronized transcription converting vocal evidence into written legal language. AI also assists in judicial decision-making by extracting information from legal texts to provide frameworks for judgment generation and sentencing prediction. The "automatic reason-generation" framework helps maintain decision consistency by matching relevant laws to facts and generating "reasons for judgment." This framework aims to speed up evidence submission, transfer of case files, and enhance access to justice.[21] This advancement led to the development and deployment of of 'Wise Judge' ('Rui Fa Guan' in Chinese) system, relying on nationwide judgment data drawn from China Judgments Online.[22] In July 2017, China's State Council unveiled the 'New Generation Artificial Intelligence Development Plan' (新一代人工智能发展规划), which sets forth the country's strategy for advancing artificial intelligence (AI). This plan articulates China's ambitions to become the global leader in AI by 2030, transform AI into a trillion-yuan (approximately 150 billion dollars) industry, and take the lead in establishing ethical norms and standards for AI.[23]

IX. ETHICAL STANDARDS AND TRANSPARENCY

As discussed, transparency and fundamental rights globally are regulated through setting ethical standards for judicial AI. These standards have been developed, engaging legal scholars, ethicists, technologists, policymakers, and international bodies. Key contributors include bodies such as the European Union, United Nations, and OECD. UNESCO has also made significant contributions through its AI ethics recommendations.

They make sure that handling of judicial decisions and data must pursue well-defined objectives, strictly adhering to the fundamental rights enshrined in the respective jurisdictions of the countries developing the system as well as right of protection of personal data. When artificial intelligence instruments are employed to adjudicate disputes, assist in judicial decision-making, or provide guidance to the public,

---

[20] Official Webpage of Hangzhou Internet Court, https://www.netcourt.gov.cn/ (Last visited June 27, 2024).

[21] Celine Cousineau, *Supra note 3*.
U. Ahmed, Z. Fatima, & T. Abbas, Implementing Artificial Intelligence (AI) into the Judicial System in Europe: Challenges and Opportunities, 8 Pakistan Soc. Sci. Rev. (Jan.-Mar. 2024).

[22] In July 2020, the Supreme People's Court (SPC) issued the Guiding Opinions Concerning Strengthening Search for Similar Cases to Unify the Application of Law, aiming to guide judges in using principles derived from prior cases to fill gaps in legislation and judicial interpretations.

[23] Roberts, Huw & Cowls, Josh & Morley, Jessica & Taddeo, Mariarosaria & Wang, Vincent & Floridi, Luciano. The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation, 36 AI & Society 10 (2021), https://doi.org/10.1007/s00146-020-00992-2.

it is of paramount importance to ensure that these instruments do not erode the guarantees of the right of access to a judge and the right to a fair trial, including the principle of equality of arms and the respect for the adversarial process. These tools must be employed with full regard for the principles of the rule of law and the autonomy of judges in their adjudicative functions. Consequently, ethical-by-design or human-rights-by-design methodologies should be favoured. This entails that, from the earliest stages of design and development, measures are integrated to preclude any direct or indirect transgressions of the fundamental values.

Given the capability of the technology to identify discrimination through data classification, stakeholders must ensure these methods do not perpetuate or worsen biases, nor lead to deterministic interpretations. Special care is required during development and implementation, especially with "sensitive" data like racial or ethnic origins, socio-economic status, political views, religious beliefs, trade union memberships, genetic and biometric data, health information, or sexual orientation. Upon detecting discrimination, corrective actions should be taken to mitigate or neutralize these risks, alongside raising stakeholder awareness.

Designers of machine learning models should leverage the expertise of relevant justice system professionals, such as judges, prosecutors, lawyers, and scholars in law and social sciences, including economists, sociologists, and philosophers. Data derived from judicial decisions and entered into machine learning software should come from certified sources and remain unmodified until used by the learning mechanism, ensuring the process is traceable and the content and meaning of the decisions are preserved.

One approach is to ensure complete technical transparency, such as through open-source code and documentation, however, this may be limited by trade secret protections. Independent authorities or experts could be responsible for certifying and auditing these processing methods or providing preliminary advice. Public authorities could issue certifications, subject to regular review. When implementing any artificial intelligence-based information system, it is also essential to provide computer literacy programs for users and facilitate discussions involving justice system professionals.[24]

X. PRACTICAL CONSIDERATIONS FOR JUDICIAL USE OF AI

Judicial office holders also must pay close attention to certain issues while using AI in their routine work. They should not enter private or confidential information into public AI chatboat. AI chatboats remember inputs and can use them for future responses. Chat history in AI chatboats is to be disabled when possible. Permissions for AI apps that request access to device information are to be refused. Verify the accuracy of AI-generated information before using it. They need to be Aware of Bias. Work devices and work email addresses should be used for AI tools.[25]

While AI's ability to process information at high speed is invaluable, it is again the judge who must bring to bear legal wisdom, experience, and the principles of justice

[24] European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment, Adopted at the 31st Plenary Meeting of the CEPEJ, Strasbourg, 3-4 Dec. 2018.
[25] Artificial Intelligence (AI) – Judicial Guidance, Dec. 12, 2023, www.judiciary.uk. (Last visited May 28, 2024).

to interpret AI's findings within the broader context of the judicial process. While AI offers profound assistance to judges in various aspects of their roles, it is paramount that this technology is applied with discernment. AI's assistance should be harnessed to enhance judicial efficiency, consistency, and fairness, without undermining the indispensable human elements of empathy, moral judgment, and ethical considerations intrinsic to the justice system.

XI. STRATEGIC ROADMAP FOR TRANSFORMING PAKISTAN'S JUDICIAL SYSTEM INSPIRED BY THE CHINESE MODEL OF SMART COURTS

Phase 1: Foundational Assessment and Strategic Planning

1. **Comprehensive Assessment**:

- Conduct a thorough evaluation of the current judicial infrastructure to identify critical bottlenecks, such as case backlogs and resource deficiencies.

- Formulate a strategic plan detailing the integration of advanced technologies, setting measurable objectives, key performance indicators (KPIs), and clear timelines.

2. **Digitization of Judicial Records**:

- Initiate a nationwide project to digitize all existing court records, transitioning from paper-based to electronic systems.

- Establish a centralized, secure database accessible to all judicial entities to streamline record-keeping and data management.

3. **Infrastructure Development**:

- Invest in robust IT infrastructure, including high-speed internet, secure servers, and cutting-edge software solutions.

- Implement comprehensive training programs to equip judicial staff and court personnel with the necessary digital skills.

Phase 2: Technology Integration in Judicial Processes

1. **Implementation of E-Filing Systems**:

- Deploy a user-friendly electronic filing (e-filing) system enabling litigants to submit documents and cases online, reducing physical paperwork and improving efficiency.

- Ensure the system is intuitive and accessible to all users, including those with limited technical expertise.

2. **Advanced Case Management Systems**:

- To enhance the efficiency and transparency of the judicial process, it is proposed to develop and integrate a comprehensive online case management system. This system should be designed to monitor the progress of cases, schedule hearings, and manage court dockets effectively. While the electronic case management system is currently operational in major cities, there is a pressing need to extend its implementation nationwide. This expansion will ensure uniform access to judicial services, streamline case handling, and improve the overall effectiveness of the legal system across the Country.

- Incorporate AI tools to automate routine tasks such as document classification and case prioritization, enhancing operational efficiency.

3. **Pilot Testing and Feedback Loops**:

- Roll out pilot programs in select courts to test and refine new technologies and workflows.

- Collect and analyze feedback from judges, attorneys, and litigants to optimize the system before full-scale implementation.

Phase 3: Enhancing Judicial Transparency and Efficiency

1. **Livestreaming Judicial Proceedings**:

- Introduce livestreaming capabilities for court hearings to promote transparency and public accountability in the judicial process.

- Develop and enforce guidelines to protect the privacy and rights of all participants in the judicial proceedings.

2. **AI-Driven Judicial Tools**:

- Implement AI-powered voice-to-text transcription services to ensure accurate and timely documentation of court proceedings.

- Utilize AI for advanced legal research, enabling rapid access to relevant case law and legal precedents.

3. **Intelligent Case Management**:

- Deploy intelligent case management systems that leverage AI to predict case outcomes, assist in decision-making, and maintain consistency in judicial rulings.

- Ensure continuous data updates and system transparency to maintain trust in AI-assisted judicial processes.

Phase 4: Scaling and Integrating Advanced Technologies

1. **National E-Evidence Platform**:

- Establish a blockchain-based national e-evidence platform to authenticate and securely exchange digital evidence.
- Seamlessly integrate this platform with existing court management systems to ensure interoperability and efficiency.

2. **Ongoing Training and Professional Development**:

- Implement continuous training programs for judges, lawyers, and court staff to keep abreast of technological advancements and their judicial applications.
- Foster a culture of continuous learning and adaptation to technological innovations.

3. **Public Engagement and Accessibility**:

- Launch awareness campaigns to educate the public about new digital judicial services and their benefits.
- Ensure these services are inclusive, providing necessary support for vulnerable populations to access justice.

Phase 5: Continuous Improvement and Innovation

1. **Systematic Feedback and Monitoring**:

- Establish robust mechanisms for collecting and analyzing feedback from all stakeholders to identify areas for continuous improvement.
- Regularly monitor the impact of technological innovations on judicial efficiency and justice delivery.

2. **Encouraging Innovation**:

- Stay informed about global advancements in legal technology and explore their potential applications in Pakistan's judicial context.
- Scale successful initiatives across all courts, ensuring uniformity and standardization.

3. **Ethical and Legal Governance**:

- Develop and enforce comprehensive ethical guidelines and legal frameworks governing the use of AI and technology in the judiciary.
- Uphold the principles of fairness, transparency, and justice in all technological implementations.

CONCLUSION

By adhering to this strategic phased roadmap, Pakistan can effectively transform its judicial system into a modern, efficient, and transparent institution,

drawing on the successful practices employed by China. This transformation will enhance access to justice, improve public trust, and ensure the judiciary's alignment with contemporary technological standards.

HURTFUL DIGITAL COMMUNICATIONS IN HONG KONG: STAKEHOLDER ACCOUNTS

Stuart Hargreaves*

Abstract: Hurtful digital communications (HDC) is an umbrella term referring to a wide variety of text and image-based practices online, such as cyberbullying or the non-consensual distribution of intimate images (NCDII). They can cause significant harm to victims and, as elsewhere, these practices are on the rise in Hong Kong. This paper represents the first stage of a project aimed at developing a coherent, broad-spectrum response to that rise. It discusses nine interviews with stakeholder groups or representatives in Hong Kong, revealing commonalities and themes in their experiences with clients who have been victims of HDC. While Hong Kong has recently adopted piecemeal reform of criminal laws targeting certain HDC practices such as 'up-skirting', NCDII, and doxing, these interviews suggest that these reforms alone are unlikely prove effective. Amongst other things, the interviews reveal a desire by victims not simply for punishment but resolution and restitution. Other jurisdictions have responded to these desires through the adoption both of expanded civil actions and broader regulatory regimes. This paper sets the groundwork for justifying similar reforms in Hong Kong.

Keywords: Law; Technology; Cyberbullying; NCDII; Communications; Hong Kong; Privacy; Tort; Regulation

* The Chinese University of Hong Kong, Faculty of Law, China.

Table of Contents

Introduction 125

I. Hurtful Digital Communications (HDC) 126

II. Hong Kong's Existing Legal Framework 127

III. Stakeholder Interviews 129

A. Interview 1 130

B. Interview 2 132

C. Interview 3 133

D. Interview 4 134

E. Interview 5 136

F. Interview 6 138

G. Interview 7 139

H. Interview 8 140

I. Interview 9 141

IV. Analysis 142

Conclusion 144

INTRODUCTION

Modern societies are increasingly composed of 'digital natives' who conduct substantial portions of their lives online from a young age.[1] Unfortunately in so doing many experience abusive behaviour from others they encounter online including threats, harassment, defamation, bullying, the non-consensual distribution of intimate images (NCDII), and doxing.[2] These are all examples of what this paper terms 'hurtful digital communications' (HDC). Regrettably, various surveys indicate such conduct is not rare in Hong Kong. Nearly one in five respondents to a 2021 Equal Opportunities Commission survey said they had been sexually harassed online within the previous year.[3] More than half of those said they had been sent unwanted sexual material; nearly half said they had directly been sent unwanted sexually suggestive messages online; just over a quarter reported being the victims of unwanted sexually suggestive comments or jokes online.[4] A separate survey of teenagers in Hong Kong in 2022 found that 40% been exposed to unwanted sexual material or messages online, while 20% reported being cyberbullied.[5] In another 17% of secondary school students reported being cyberbullied.[6] In 2023, a local non-governmental organization (NGO) revealed that it had assisted 646 victims of NCDII over a two-year period.[7]

There is also research that focuses on the perpetrators. In one study, nearly one in three youths had admitted to perpetrating HDC themselves, while one in four respondents said they had altered images and posted them online to mock or embarrass someone.[8] A 2018 survey by researchers at the University of Hong Kong suggested that 2.6% of students age 13-17 had attempted to dox someone.[9] Rates skyrocketed the following year when doxing of perceived ideological opponents became part of the social unrest in Hong Kong,[10] with the Office of the Privacy Commissioner for

---

[1] See e.g., Mark McCrindle, *Generation Alpha* (Hachette UK, 2021).

[2] See e.g., Robert Faris et al., "Understanding Harmful Speech Online" (2016) Research Publication No. 2016-21, *Berkman Klein Center for Internet & Society at Harvard University*, online: https://cyber.harvard.edu/publications/2016/UnderstandingHarmfulSpeech.

[3] "Survey on Sexual Harassment in Hong Kong 2021", *Equal Opportunities Commission*, May 2022, online: https://www.eoc.org.hk/compass/wp-content/uploads/2022/05/Territory-wide-Representative-Survey-on-SH-in-HK-2021-Infographic-EN.pdf.

[4] Ibid.

[5] "Hong Kong Kids Online", *Save the Children*, May 2022, online: https://savethechildren.org.hk/wp-content/uploads/2022/05/Hong-Kong-Kids-Online-Report-English-Final.pdf.

[6] "Survey Results on Cyberbullying in Hong Kong", *Zonta Club of Hong Kong East*, 14 Jan 2022, online: https://m21.hk/zontaHKE2023/download/ZONTAHKE_Cyberbully_Survey_Report_2022.pdf

[7] Irene Chan, "Hong Kong NGO requests removal of over 1,300 non-consensual intimate images from online platforms" *Hong Kong Free Press*, 8 Aug 2023, online: https://hongkongfp.com/2023/08/08/hong-kong-ngo-alerted-to-over-1300-intimate-images-online-as-it-urges-platforms-to-honour-removal-requests/.

[8] Emily Hung, "How common is cyberbullying among young people in Hong Kong? NGO says third are targeting peers online and calls for more prevention measures", *South China Morning Post*, 9 Sep. 2023, online: https://www.scmp.com/news/hong-kong/education/article/3233991/how-common-cyberbullying-among-young-people-hong-kong-ngo-says-third-are-targeting-peers-online-and

[9] Ruby Fung & Michael Cheung, "An in-depth case study of doxing in Hong Kong", *Tackling Cyberbullying: A Comparative and Interdisciplinary Symposium*, 4 Sep. 2018, online: https://www.hku.hk/f/upload/18335/Ruby%20Lo.pdf.

[10] See Anne Cheung, "Doxing and the Challenge to Legal Regulation: When Personal Data Becomes a Weapon" in Jane Bailey, Asher Flynn & Nicola Henry, eds., THE EMERALD INTERNATIONAL HANDBOOK OF TECHNOLOGY-FACILITATED VIOLENCE AND ABUSE (Emerald Publishing Limited, 2021) 577; see also Yao-Tai Li & Katherine Whitworth, "Coordinating and doxing data: Hong Kong

Personal Data (PCPD) reporting a nearly 400% increase cases between 2018 and 2019.[11] While the survey results may differ year on year and as between different organizations and methodology, the above information shows that in a general sense – as in other jurisdictions – the frequency of HDC in its various forms appears to be on an upward trend in Hong Kong.

This paper describes the first stage of a multi-year project that seeks to identify appropriate *non-criminal* forms of regulating HDC conduct. This stage consisted of a series of interviews with local stakeholder groups and individuals; generally speaking non-governmental organizations in Hong Kong who work with or advocate for victims of HDC rather than victims themselves. It collates and analyses the commonalities and themes emerging from the interviews that ought to inform future legislative and regulatory change. The interviews suggest that the problem of HDC is on the rise in Hong Kong and affects even young children, that cyberbullying and NCDII are the two most common forms of HDC that relevant NGOs come across, a dissatisfaction with how the relevant criminal law is applied, and that a lacuna in the law that makes obtaining removal of harmful material online more challenging than it ought to be.

I. HURTFUL DIGITAL COMMUNICATIONS (HDC)

Lenhart et al. argue that the shared element to various HDC practices is "unwanted contact that is used to create an intimidating, annoying, frightening, or even hostile environment for the victim and that uses digital means to reach the victim."[12] Notwithstanding this digital mediation, HDC can result in severe consequences for victims including anxiety, embarrassment, shame, depression, social withdrawal, and even thoughts of suicide.[13] There can be reputational harm,[14] harassment by third parties,[15] and professional/financial loss as victims withdraw from social or economic life.[16] The digital nature of HDC means that victimisation can be an ongoing state –

protesters' and government supporters' data strategies in the age of datafication" (2023) SOCIAL MOVEMENT STUDIES 1–18. https://doi.org/10.1080/14742837.2023.2178404.

[11] "PCPD Sets Up Enquiry/Complaint Hotline about Doxxing and Releases the Results of a Survey on Protection of Personal Data Privacy", *Office of the Privacy Commissioner for Personal Data*, 28 Jan. 2021, online: https://www.pcpd.org.hk/english/news_events/media_statements/press_20210128.html.

[12] Amanda Lenhart, Michelle Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney, "Online Harassment, Digital Abuse, and Cyberstalking in America" *Data & Society Institute*, November 21, 2016, online: https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/, cited in Robert Faris et al., *supra* note 3.

[13] See e.g., Laura Leets & Howard Giles, "Words as weapons - when do they wound? Investigations of Harmful Speech" (1997) 24 (2) HUMAN COMMUNICATION RESEARCH 260; Allisdair A. Gillespie, "Cyber-bullying and the harassment of teenagers: the Legal Response" (2006) 28 (2) JOURNAL OF SOCIAL WELFARE & FAMILY LAW 123; Robert Faris et al., *supra* note 3; Samantha Bates, "Revenge Porn and Mental Health: A Qualitative Analysis on the Mental Health Effects of Revenge Porn on Female Survivors" (2017) 12(1) FEMINIST CRIMINOLOGY 22; Penza, Dylan E., "The Unstoppable Intrusion: the Unique Effect of Online Harassment and What the United States Can Ascertain from Other Countries' Attempts to Prevent It" (2018) 51(1) CORNELL INTERNATIONAL LAW JOURNAL 297.

[14] See e.g., Danielle Keats Citron & Mary Anne Franks, "The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform" (2020) UNIVERSITY OF CHICAGO LEGAL FORUM 45; Ganaele Langlois & Andrea Slane, "Economics of Reputation: the Case of Revenge Porn" (2017) 14(2) COMMUNICATION AND CRITICAL/CULTURAL STUDIES 120.

[15] See e.g., Nicola Henry, Asher Flynn & Anastasia Powell, "Responding to 'revenge pornography': Prevalence, nature and impacts" (2019) *Report to the Criminology Research Advisory Council* (CRG 08/15-16).

[16] See e.g., Citron & Franks, *supra* note 15.

once an image, video, or abusive message is uploaded to a website, it may be virtually impossible to prevent its further spread, regardless of criminal penalty.[17]

II. HONG KONG'S EXISTING LEGAL FRAMEWORK

One way to tackle the rise of HDC would be to adopt a 'broad spectrum' approach, relying on a combination of criminal prohibition, civil penalties, and regulatory bodies tasked with responding to the problem. However, relevant law reform in Hong Kong has thus far not adopted such a comprehensive and integrated approach. Instead, there have been piecemeal changes that have focused on criminal prohibition of particular activities after they reach a certain level of public salience.

After many years of discussion,[18] in 2021 the Crimes Ordinance was amended to create four new offences related in part to HDC, each with a maximum penalty of five years' imprisonment.[19] The new offences are voyeurism, unlawful recording/observation of intimate body parts, publication of matters arising from those two offences, and publication or threatened publication of intimate images without the subject's consent.[20] The amendments also allow magistrates to issue 'disposal orders' at any time during the proceedings, ordering any person in Hong Kong or elsewhere to take steps to remove, delete, or destroy an image that is the subject of a proceeding relating to one of the four offences.[21]

The Personal Data (Privacy) Ordinance (PDPO)[22] was also amended in 2021 to create two new offences[23] in response to the dramatic increase in doxing incidents referred to above. A summary offence proscribes the disclosure of personal data without consent and with intent or being reckless as to whether a specified harm will occur to either the data subject or their family;[24] the second, indictable offence is satisfied where a specified harm actually occurs.[25] The specified harms are harassment, molestation, pestering, threat, intimidation, bodily or psychological harm, causing someone to fear for their physical safety, or damage to personal property.[26] The summary offence is

---

[17] See e.g., Samantha Kopf, "Avenging Revenge Porn" (2014) 9 (2) MODERN AMERICAN 22; Olga Marques, "Intimate Image Dissemination and Consent in a Digital Age: Perspectives from the Front Line" in Bailey, Flynn, & Henry, eds., *supra* note 11.

[18] See e.g., "Report on Voyeurism & Non-Consensual Upskirt Photography", *The Law Reform Commission of Hong Kong*, Apr 2019, online: https://www.hkreform.gov.hk/en/docs/rvoyeurism_e.pdf; see also Thomas Crofts, "Criminalization of Voyeurism and 'Upskirt Photography' in Hong Kong: The Need for a Coherent Approach to Image-Based Abuse" (2020) 8(3) THE CHINESE JOURNAL OF COMPARATIVE LAW 505.

[19] Crimes (Amendment) Ordinance 2021, Ord. No. 35 of 2021, A3597, online: https://www.elegislation.gov.hk/hk/2021/35!en. For discussion of these offences in the context of NCDII, see Thomas Crofts, "Combatting Intimate Image Abuse in Hong Kong", (2022) 52 HONG KONG LAW JOURNAL 405.

[20] Crimes Ordinance, Cap. 200, s. 159AA (B – E).

[21] Ibid., s. 159AA (L).

[22] Personal Data (Privacy) Ordinance, Cap. 486 (PDPO).

[23] Personal Data (Privacy) (Amendment) Ordinance 2021, Ord No. 32 of 2021, A3363, online: https://www.gld.gov.hk/egazette/pdf/20212540/es12021254032.pdf.

[24] PDPO, s. 64(3A)

[25] PDPO, s. 64(3C)

[26] PDPO, s. 64(6).

subject to a punishment of imprisonment for two years;[27] the maximum punishment for the indictable offence is imprisonment for five years.[28]

But while these new offence provisions may provide deterrence and punishment for particular pernicious behaviours, the interviews in this article suggest the laws still do not adequately respond to the needs of victims. The criminal process is complex, time-consuming and uncertain. The police must become involved. Prosecutors must decide to proceed. Conviction requires proof beyond a reasonable doubt. The process requires victims to engage with the justice system and may require them to relive their experiences.[29] Criminal law primarily focuses on penalties for the perpetrators, rather than mechanisms of restitution for the victims – while this approach may indeed serve to deter some harms, it does little to provide redress for actual victims.

This research project began from the standpoint that civil law and administrative law mechanisms may provide further avenues to combat the pernicious rise of HDC. Some forms of HDC – in particular NCDII – involve an apparent invasion of privacy, yet the civil protection of privacy in Hong Kong remains incomplete. The Law Reform Commission recommended the introduction of a statutory privacy tort two decades ago,[30] but these reforms have not been implemented and courts have yet to adopt a freestanding tort of privacy in Hong Kong.[31] This means private law remedies for HDC are only available via traditional torts such as assault, defamation, and the action for intentional infliction of harm.

The lower Hong Kong courts have in recent years moved towards recognising a new tort of harassment that might prove useful. Under one proposed version, the tort would require a course of conduct "sufficiently repetitive in nature... to cause worry, emotional distress or annoyance."[32] Though the Court of Final Appeal has yet to rule on the final contours of such a tort, it is at least conceivable that under it some examples of HDC such as repeated cyberbullying or *ongoing* threats to distribute intimate images might fit within it. But since the traditional common law torts did not evolve in an era of widespread abusive online conduct generally speaking they are not a good fit for rendering most forms of HDC actionable. For example, under the current jurisprudence a single threat to publish or even the actual publishing of an intimate image without the subject's consent likely would not fit into any of the standard actions. Infliction of harm requires personal injury, including psychiatric injury; defamation is subject to the defence of truth; and the tort of assault requires the threat of bodily harm, etc. A subsequent stage of this project will therefore consider the potential creation of

---

[27] PDPO, s. 64(3B).
[28] PDPO, s. 64(3D).
[29] Tyrone Kirchengast & Thomas Crofts, "The legal and policy contexts of "revenge porn" criminalisation: the need for multiple approaches" (2019) 19(1) OXFORD UNIVERSITY COMMONWEALTH LAW JOURNAL 1; Michael Salter, Thomas Crofts, & Murray Lee, "Beyond criminalisation and responsibilisation: Sexting, gender and young people" (2013) 24(3) CURRENT ISSUES IN CRIMINAL JUSTICE 301.
[30] "Report on Civil Liability for Invasion of Privacy", *The Law Reform Commission of Hong Kong*, Dec. 2004, online: https://www.hkreform.gov.hk/en/docs/rprivacy-e.pdf.
[31] See also Jojo YC Mo & AKC Koo, "A Bolder Step towards Privacy Protection in Hong Kong: A Statutory Cause of Action" (2014) 9(1) ASIAN JOURNAL OF COMPARATIVE LAW 345.
[32] *Lau Tat Wai v Yip Lai Kuen Joey* [2013] HKCFI 369, 62, citing for the boundaries of the tort directly from the Singaporean case of *Malcomson Bertram & Anor v Naresh Mehta* [2001] 4 SLR 454, 464F.

statutory torts that might be more closely aligned with the harms suffered by victims of HDC.

The primary legislative approach to protecting privacy in Hong Kong is the PDPO. However, the PDPO is aimed at *informational* privacy only. That is, the privacy of individuals as it pertains to their personal data rather than privacy more broadly. Beyond the aforementioned creation of offences related to doxing, the PDPO therefore has a relatively limited role to play in regulating or controlling HDC. The PDPO's six 'data protection principles' are concerned with matters such as the collection, accuracy, use and retention of personal information. While in theory the law might be engaged by HDC that involves the handling of personal information, thanks to an early decision the PDPO offers less protection in such a situation than one might first expect. In *Eastweek*, the Hong Kong Court of Appeal ruled that a bare photograph – *even if* it shows an identifiable face – does not meet the definition of 'personal data' within the meaning of the PDPO (that the information is "about an identified person or about a person whom the data user intends or seeks to identify."[33] On this basis, a majority of the Court denied that it amounted to a collection of personal data for a photographer to take photos of unknown individuals in a public place and for a news magazine to publish these photographs. It appeared not to make a difference to the majority that the publication of the photograph allowed third parties, such as readers of the magazine, to identify the subject of the photograph.[34] Unfortunately, it appears that the PCPD continues to rely on *Eastweek* to turn away victims of NCDII requesting assistance on the ground that the publication of photographs showing individuals is not a breach of the PDPO unless the material is accompanied by other personal information.[35] Moreover, even if the courts revisited this position, breach of the data protection principles within the PDPO is not an offence in and of itself.[36]

These brief observations are sufficient to demonstrate some of the weaknesses in the current legal regime and regulatory response to HDC. Our project sought to bolster this initial analysis with stakeholder accounts. In its first stage, we identified and interviewed various stakeholders in the community with an interest or expertise in the area of hurtful digital communications in an attempt to understand the state of HDC 'on the ground' in Hong Kong. The next part of the paper collates and documents the results of these interviews.

III. STAKEHOLDER INTERVIEWS

---

[33] *Eastweek v Privacy Commissioner for Personal Data*, [2000] HKCA 442. Writing for the majority, Ribeiro JA argued 'What is crucial here is the complainant's anonymity and the irrelevance of her identity so far as the photographer, the reporter and Eastweek [the magazine] were concerned. Indeed, they remained completely indifferent to and ignorant of her identity right up to and after publication of the offending issue of the magazine' (at para. 14).

[34] In contrast, Wong JA's dissent in *Eastweek* acknowledged that it is quite easy for someone to emerge from anonymity when a photograph of them is widely published even without accompanying information, once one person recognizes them (at para. 46).

[35] See "Privacy Commissioner urged to tackle revenge porn", *RTHK*, 8 Aug. 2023, online: https://news.rthk.hk/rthk/en/component/k2/1712588-20230808.htm.

[36] See Jojo YC Mo, "Are data protection laws sufficient for privacy intrusions? The case in Hong Kong", (2014) 30(4) COMPUTER LAW & SECURITY REVIEW 429.

Stakeholders we interviewed included NGOs working with victims of HDC and two legal representatives who had performed pro bono services for victims.[37] We decided at the outset of the project to not interview victims directly, for several reasons. First, we considered that given the frequently sexualized nature of the material at issue, victims might be uncomfortable discussing their experiences with us. We were keen to avoid anything that might cause a 're-victimisation' experience, and so decided speaking instead with representative stakeholders was the more appropriate route. Second, by focusing our limited time and resources on stakeholder organizations and bodies, we were able to gather a wider range of experiences than we would have had we focused primarily on individual victim accounts.

We conducted interviews with each of these stakeholders in person, and recorded them (audio only) after obtaining participants' written consent; we also obtained ethics approval for the interviews as part of the grant-seeking process.[38] With two exceptions, the interviews were conducted in English. A transcript of each interview was made by our research assistant, who was also responsible for translating two interviews in which we asked questions in English but the interviewees responded in Cantonese. When obtaining the written consent of the interviewees to record and transcribe the conversations, we agreed to not include their names or those of their organizations in our research outputs since they are often dependent at least in part on public funding. Thus, the summaries below refer in general terms to the kind of organization the interviewee represented.

A. Interview 1[39]

This was an interview with a representative of an NGO that offers a range of services to sex workers in Hong Kong, including outreach and counselling provided online (via Twitter/X, Instagram, and Heymandi) and by 24-hour telephone hotline. Clients who reach out are offered health checks and the face-to-face interaction stemming from that is often the springboard for discovering whether they are in crisis. Initially the organization primarily dealt with issues arising from compensated dating[40] amongst under 18s, unwanted pregnancies, and providing advice to clients on what to do in cases of undercover police operations. The interviewee relayed however that in recent years an increasing number of their clients had fallen victim to secret photography or video recording leading to threats and blackmail. Without being able to provide formal statistics, they estimated that roughly 30% of their clients complained about experiencing behaviour that we classed as a hurtful digital communication.

---

[37] We also conducted one 'off the record' interview with a Government representative who provided background information on the application of the relevant criminal laws.

[38] Approval given by the Survey & Behavioural Research Ethics Committee of the Chinese University of Hong Kong, application SBRE-21-100, 29 Dec. 2021.

[39] Interview conducted 9 May 2013. Questions asked in English, answers provided in Cantonese and then translated and transcribed by the research assistant. Transcription on file with the authors.

[40] The phrase refers to the practice where youth (primarily but not exclusively female) enter into relationships that may contain both sexual and non-sexual components in exchange for a range of economic benefits. The frequent non-sexual aspects of the relationships tend to distinguish the practice from more conventional prostitution, though compensated dating can often lead to prostitution over time. For discussion of the practice in Hong Kong, see e.g., Tak Yan Lee & D.T.L. Shek, "Compensated dating in Hong Kong: prevalence, psychosocial correlates, and relationships with other risky behaviors" (2013) 26(3) JOURNAL OF PEDIATRIC AND ADOLESCENT GYNECOLOGY S42-S48.

The interviewee (again anecdotally) reported that the majority of their clients they dealt with were teenagers, but they had clients up to age 25 and as young as 12. They also felt that their clients were getting younger, in part because during COVID19 more young students spent time online rather than attending school. When schools in Hong Kong closed during the pandemic, said the interviewee, many students lost access to social workers provided through those institutions. Students had more free time, a need for money, and in the interviewee's view this had led to a spike in young girls entering into 'compensated dating' relationships. The interviewee reported that while their clients were typically careful to not post images of their face or information about their real identities when seeking clients, sometimes the men interested in procuring their services would request they verify their identity during online conversations. This increased the risk of identification since they would feel pressured to agree to the request. The interviewee also reported that their clients found that during conversations or when meeting in real life, men would often try to obtain personal information (such as which school they went to) that was then often later used for blackmail.

The interviewee had on one occasion assisted a client in making take down requests to a website to which sexually explicit material in which they were visible had been posted. In that instance an (unnamed) website hosted outside Hong Kong had responded promptly and helpfully and removed the material when requested, however at that point the video had already spread within the client's school. The client ended up leaving both the school and Hong Kong. While this was their only experience requesting take-down of a sexual video or image, the interviewee expressed the view that there were many unreported cases.

The interviewee had more often accompanied a client to make a formal complaint to the police, however they reported that in practice less than 1 in 10 wanted to pursue charges. They believed this hesitation was for a variety of reasons. First, the interviewee reported that clients were generally unfamiliar with the law (in particular the recent changes that made it an offence to threaten to distribute images), and so were often not even sure if the perpetrators had done anything illegal. The interviewee also reported that perpetrators (especially those seeking sexual favours rather than money) appeared to be adjusting their behaviours in light of changes to the law – rather than making direct threats to spread an image, instead they might send a picture and just remark "this is a good photo, isn't it?" or make a relatively vague suggestion such as "you know what I want". Second, the interviewee's clients were dubious as to the probability of a successful prosecution, given there was often no clear evidence in the video of the identity of the perpetrator. Third, their clients expected that as sex workers they would receive unsatisfactory treatment and attitudes from the police. Those under 18 were particularly afraid the police would inform their families, who often would be unaware that they were working in the sex industry. As a result, in cases that did not involve physical violence the victims were rarely inclined to involve the police.

The interviewee was of the view that cultural changes were necessary in combating many of these issues – for instance, a young girl filmed without her consent should not be so afraid that the police or others would judge her for being a sex worker that she would avoid reporting the crime. Until such a larger cultural change could occur, they argued that procedural changes might assist – for example, that making sure the victims' identity was not revealed to any unnecessary individuals as part of the process of making a complaint. In terms of education, they believed that 'one-off' training

sessions in schools about cyberbullying or related issues were insufficient. Instead, continuous public education was required. In terms of regulatory bodies, the interviewee was of the view that a centralized take-down service for intimate images would be helpful. Though lacking personal experience, they also held the view that it would be easier to obtain removal from a large foreign website than a smaller local forum.

B. Interview 2[41]

This was an interview of a representative of an NGO that worked largely on behalf of the 'men who have sex with men' (MSM)[42] community in Hong Kong. The interviewee reported that sometimes in their work their organization came across clients who were being blackmailed with threats to post sexually oriented images and videos of them on the internet. They described two common and one less common scenarios. In the first group of cases, nude images or sexual videos were spread online by ex-partners of their clients after a relationship had ended. The second group involved situations in which a client had sent a nude photo to someone they met on an online dating website, and then the recipient of that photo used it to try to blackmail them into having sex or paying money by threatening to send the images to the victim's friends or family. In many cases the victims were not publicly 'out', and so victims often feared consequences beyond embarrassment. A less frequent occurrence involved convincing victims to install an app on their phone (usually under the pretence that it would give them access to sexual content) which would allow the perpetrator to take control of the device. The list of the victim's contacts on the phone would then be used for further blackmail. Victims of these three scenarios ranged from age 14 to 50.

The interviewee's NGO runs a 24-hour helpline and offers information through its Instagram account on how gay men and sex workers can protect themselves online, and what to do if they are being threatened. The NGO had some experience with NCDII, and had created guides for clients seeking to obtain removal of material. In their experience, while in the past Twitter/X had been relatively responsive to requests they made on behalf of clients, since a change in ownership[43] that was no longer the case. Meta (owner of both Facebook and Instagram) and Google were more responsive to requests, but this NGO found it was still challenging for people to prove that the images were of them, since often their faces would be obscured. The interviewee had comparatively little experience dealing with local websites, which was in their view because they were "less friendly" places from the perspective of Hong Kong's gay community than were Twitter/X and Instagram. In any event, this interviewee said the NGO's approach in cases of NCDII was usually to tell clients that if an image was already being circulated online the chances of being able to completely remove it were very low, and that giving in to blackmail would likely not succeed and the demands would just increase. They instead recommended an approach based on "risk management" – for example, suggesting that the victim should consider getting in

---

[41] Interview conducted 9 May 2023. Transcript on file with authors.

[42] MSM includes all men who engage in sexual activity with other men, regardless of their sexual orientation. See e.g., "Cisgender men who have sex with men (MSM)", International Association of Providers of AIDS care, undated, online: https://www.iapac.org/fact-sheet/cisgender-men-who-have-sex-with-men-msm/.

[43] Kate Conger & Lauren Hirsch, "Elon Musk Completes $44 Billion Deal to Own Twitter", 27 Oct. 2022, *The New York Times*, online: https://www.nytimes.com/2022/10/27/technology/elon-musk-twitter-deal-complete.html.

contact with family or their workplace proactively and describe in general terms what has happened and they should not open or view files they have been sent.

The interviewee reported that despite the aforementioned changes to the Crimes Ordinance related to NCDII, in their experience the police were often not very helpful unless there were threats involving physical harm or demands for money. They believed that most of their clients wished to avoid the police due to embarrassment, or because they were afraid that the police would blame them for having sent someone a sexual image or video in the first place. They recommended more training for the police, arguing that the frontline officers who received HDC-related complaints needed to be more sensitive to the context and avoid slipping into victim blaming. They also noted that while they would try to meet with the Crime Prevention Bureau once every few months to discuss strategies, sometimes they would receive no response.

C. Interview 3[44]

This was an interview with a representative of an NGO working on sexual violence issues faced by women, including NCDII. They reported that despite the recent changes in the law, their clients remained reluctant to approach the police for at least three reasons. First, from the interviewee's perspective their clients' primary interest was in having the material removed from the internet, and the clients did not believe the police were in a position to help them accomplish this. Second, the interviewee thought it relevant that while in-person 'physical' forms of sexual violence were usually perpetrated by an intimate partner, in cases of NCDII they estimated that roughly 1/3 of their clients had no idea who the perpetrator was. If the perpetrator was not a known individual based in Hong Kong, then the chances of successful prosecution would be very low, and in such cases the whole process would be seen as a waste of time by both the victim and the police. The third reason the interviewee believed their clients were reluctant to approach the police was that they were often worried about the handling of their complaint. The interviewee shared an example in which a victim complained about the non-consensual spread of her intimate images and videos, and then those materials spread within the police force itself. The interviewee said that, to its knowledge, there were no clear guidelines for frontline officers to follow when receiving NCDII-related or similar complaints.

The interviewee reported that they were successful in having material removed from online platforms on behalf of their clients in about 85% of cases. In the remaining 15% material remained online either because it was deemed not to violate the websites' policies or because the requests were simply ignored. In the interviewee's experience major commercial/mainstream social media websites hosted overseas (eg Meta, Twitter/X – the latter at least historically) were usually very responsive because they already had specific terms of use in place regarding nudity and non-consensually shared images. In contrast, pornographic website owners or messaging apps like Telegram were far less likely to respond positively to their take-down requests. In the case of the former requests were often simply ignored, while in the latter the encrypted nature of the messaging channels often meant the host itself could not access what those channels were sharing. However, the interviewee reported that in cases where the website owners itself did not reply, sometimes they were able to contact that commercial webhost

[44] Interview conducted 17 May 2023, via zoom. Transcript on file with authors.

provider directly (for instance, Cloudflare) and have the *entire website* removed as being in violation of the host's own terms of service.

The interviewee considered that the new anti-doxxing provisions in Hong Kong's law had limited utility for their clients. As explained above, the new PDPO amendments were only relevant if the images at issue were shared alongside information about the victim like their name, address, or HKID number and the interviewee said this typically was not the case in situations of NCDII. However, they said that sometimes the PCPD would send a take-down request to a local website on their behalf even where the strict conditions was not met, hoping that the website would remove the material even absent legal compulsion. Overall the interviewee was generally dissatisfied with the current legal framework, noting that even with the changes to the Crimes Ordinance NCDII is not classed as a sexual offence. This means that any proceedings are held in open court, and victims are not able to deliver testimony via CCTV etc. The interviewee was of the view that classing NCDII as a sexual offence would provide victims with more protection and increase the likelihood of them being willing to pursue legal resolution. The interviewee said, however, that there were divided views within their organization about whether a government body like the PCPD should have the power to directly order take-down of material absent a court order, in part because of the political implications that would surround any such power.

This interviewee's organization also ran a variety of education campaigns and believed them to be an important component of the fight against HDC, and the interviewee noted that discussions about behaviour online and its connection to sexual violence was generally absent from most sex education programmes in Hong Kong schools. Their organization sought to fill this gap by providing workshops both online and in person. However, the interviewee reported that though their organization had been invited into some schools, generally they were only given a single two-hour timeslot and in their view this was inadequate.

Overall the interviewee felt that their organization had sufficient resources to deal with the numbers of take-down requests of NCDII they were receiving from their clients. At the same time they felt that the numbers of complaints were limited in part due to local cultural pressures that dissuaded people from seeking help regarding matters related to sex and sexuality in the first place; they believed there were many unreported cases.

D. Interview 4[45]

This was an interview with a representative of an NGO that provides social work services for children and youth. The interviewee described three services their NGO offered that centred around a health digital lifestyle – direct counselling of students suffering from internet addiction or HDC issues; collaborating with student groups in both schools and universities to increase awareness about the risks of certain behaviour online; and running campaigns online that focused on the same. HDC issues they had come across included cyberbullying, doxing, NCDII, and using the images of the NGO's clients to create fake accounts on apps or websites used for finding casual sex partners. The interviewee reported their clients were both male and female, and usually were in their teens – although they had one client who was only nine years old. They

[45] Interview conducted 22 May 2023. Transcript on file with authors.

reported that young boys were typically threatened with blackmail by men they met online and to whom they had sent nude images. Often times the victims were so ashamed of this they would initially pay the blackmail, only to discover that the amount demanded would simply keep rising. Only then might they confide in a social worker or teacher.

While sometimes victims would approach them directly, this NGO also was sometimes brought in by schools to offer counselling in the event of an incident. Interestingly, they reported provided counselling not only to victims, but also the perpetrators and third parties as well. For instance, in one case a student had been suspended for a hurtful digital communication involving sexual imagery of one of his classmates. That student was then given counselling to help them process the punishment they had received, but the other students *also* received training or information about how to appropriately react and deal with that perpetrator once their suspension was completed. So there was an aim at reintegration of the perpetrator into the classroom, as well as the more conventional efforts to provide for the needs of the victim.

While sometimes they were also contacted by parents worried about their children being bullied online or suffering from internet addiction, this interviewee reported they found it challenging to get some parents to voluntarily come to talks about cyberbullying. In their view, parents who were disengaged from their child's life would benefit most from such a talk and yet those were the parents who were least likely to attend. They also reported on occasion there would be conflict between school administration and parents about how to resolve an incident and they found it challenging to operate in such cases an environment.

This interviewee believed when it came to counselling, their clients who had suffered an HDC-related harm primarily wanted somebody to talk to who was supportive and open-minded. While their NGO would always offer to send a representative accompany a client to make a report to the police if there was a safety concern or if a criminal offence appeared to have been committed, the interviewee said that generally clients were reluctant to go to the police. This was because, in the interviewee's view, their clients did not expect to receive appropriate emotional support or empathy from the police if they chose to report an incident. In the interviewee's experience, victims who *did* choose to make a report to the police received a more responsive outcome when they were accompanied by a representative of their NGO. However, they also reported that there was significant variation in the approach or attitude of the police, even when they accompanied the victim. They gave one example of a client whom they accompanied to make a police report, a 14 year old girl who had voluntarily created sexual imagery with a partner. After the relationship ended that partner threatened to spread the images online if the client did not resume their relationship. The interviewee described the first officer they dealt with as helpful, respectful, and willing to investigate. However, because the alleged perpetrator was living in a different district, the file was transferred to another office in that district. The second officer was, in the view of the interviewee, dismissive and blamed the girl for having taken the photos in the first place. The interviewee believed this difference reflected the fact that not every individual district police station has a member of the cybercrime unit.

The interviewee reported mixed experiences when attempting 'take-down' of NCDII material on behalf of their clients. Their NGO was classed as a "trusted partner" on YouTube, meaning reports made by the NGO's staff would receive priority consideration and were generally successful.[46] But this was not the case with other services. The interviewee relayed that a client of theirs who had directly approached Meta was unsuccessful with a request for removal of material from Facebook because her face was obscured and she therefore could not prove that she was the person featured in the imagery. The interviewee also reported no success when approaching services like Telegram or local websites to request removal of imagery of their clients.

The interviewee was of the view that it would be helpful to have stricter laws on HDC, if only for the deterrent effect. At the same time, they believed that law reform alone was insufficient and that continued preventative education campaigns were necessary, since digital anonymity meant that many activities might remain largely out of reach of the authorities.

E. Interview 5[47]

This interview was with representatives of a large Hong Kong-based organization offering a wide range of services, including social work, education, community outreach, and medical care. The representatives we spoke with worked for two specialized departments within this organization that related to our project theme. The first department worked closely with the perpetrators and victims of sexual offences, while the second worked primarily online but with a focus on education.

The interviewee from the first department noted the department's focus had changed since it was initially setup approximately 15 years ago. While this team had originally been set up to help re-habilitate perpetrators of sexual offences, it was found that many of the perpetrators of online sexual offences (including hurtful digital communications) had themselves been victims at an earlier stage. The department had thus increased its efforts to break the cycle of violence by offering services to the victims as well. Perpetrators were generally referred to the interviewee's department by probation officers, while victims were usually referred by school workers, parents, or family centres. This department does not provide direct legal assistance to either group, but does offer basic information to individuals who have been charged with an offence about the criminal process. Where probation orders required ongoing counselling with the organization, this involved combinations of education and therapy, as well as group sessions with other youth sex offenders.

The interviewee of this department reported that the perpetrators they worked with ranged in age from 8 to 24, with a 70-30% split male and female; the victims were in the same age range, but split 50-50% along gender lines. The interviewee reported that the nature of the offences they came across had changed over the years. Initially they dealt with offences such sexual assault or unlawful intercourse, while increasingly they were confronted with activities that had some kind of digital or online component. These new issues included NCDII, sexual humiliation online, 'up-skirting', bullying

---

[46] "YouTube Partner Programme Overview and Eligibility", *YouTube*, undated, online: https://support.google.com/youtube/answer/7285; "Get in Touch with the YouTube Creator Support Team", YouTube, undated, online: https://support.google.com/youtube/answer/3545535.

[47] Interview conducted 23 May 2023. Transcript on file with the authors.

within an online game, etc. The interviewee said a particularly common problem they dealt with was minors who were being blackmailed after having been convinced to send nude photos to someone they met online.

While the second interviewee's department had originally conducted in-person district level outreach work for youth with mental health problems, in recent years it had pivoted to primarily online work, focusing largely on offering support to youth exposed to some kind of HDC. Their approach to 'digital outreach' was quite innovative, including attempts to connect with youth through popular online games. The interviewee reported that the cyberbullying they dealt with was sometimes an extension of real-world bullying, but sometimes it was a response to it. In other words, sometimes a victim of physical bullying (at school, for example) becomes the perpetrator of cyberbullying as a kind of revenge.

School campaigns conducted include 'empathy training' for students to try and help them understand the consequences of their actions online, with the idea that hurtful digital communications were not 'just jokes'. Emotional regulation is another focus, teaching students how to deal with inter-personal problems in constructive rather destructive ways. The interviewee reported they their department was happy to run education campaigns in any school that was interested, and that they believed it was most effective if done proactively, beginning at the elementary school level. They adopt a 'train the trainers' approach, educating not just children but also school staff on how to spot the signs of cyberbullying.

However, the interviewee reported that in practice their department tended to only be called in *after* a serious incident had occurred. They suggested that this reactive approach was less effective than a proactive one, as students and staff needed to be prepared *before* incidents actually happened. While not having access to numbers to prove it, this interviewee considered that the reactive posture might in part be a financial issue – they believed schools tended to have a lump sum in their budget set aside for social work programmes generally, and bringing in an outside provider (or bringing on in on a regular basis) for a specific programme might be a cost that not all could afford. The interviewee acknowledged that there were many kinds of social work programmes that a school might want to bring in, and campaigns related to HDC would be only one option among many.

The interviewee believed that a second obstacle they faced in conducting pre-emptive education campaigns was that schools might be worried that bringing them in might be interpreted by parents as a sign that there 'was a problem' – the interviewee described being asked by a school to refer to their campaign as being about 'interpersonal relationship problems' rather than 'bullying'. A final potential obstacle to effectively dealing with HDC in schools was that, in the view of this interviewee, sometimes the problem was not seen as 'serious' by school leadership. This was especially the case if the events lacked a connection to a 'real world' confrontation between students, although the interviewee also noted that a younger generation of teachers who had grown up as digital natives themselves tended to have a more nuanced perspective.

In addition to running educational campaigns regarding cyberbullying and doxing, this department also partnered with Meta[48] to allow it to quickly report posts they found appeared to be cyberbullying or that encouraged suicide. However, the interviewee perceived a lack of transparency in how Meta made the final decisions about removal of material. Even more problematic was, they said, dealing with local websites such as LiHKG. The interviewee felt that while LiHKG is a hotbed of cyberbullying and doxing in Cantonese, its administrators rarely reply to their requests to remove harmful content. The interviewee also believed that the rise of encrypted communications services (eg WhatsApp) were creating new avenues for bullying, as the service provider could not automatically scan for certain kinds of material.

In such cases, the primary remedy was to report the activities to the police if it crossed a line into threatening the victims' safety. However, the interviewee reported that often the police were reluctant to assist unless the victim provided detailed evidence (such as a printout of threats). The interviewee also believed that in most cases of HDC the police were more inclined to respond positively complaints made by their NGO *on behalf of* the victim, rather than by the victim directly. Exceptions to this pattern was where the HDC was accompanied by physical abuse or had a clear 'real-world' component such as 'up-skirting'; in the view of the interviewee, most forms of HDC were a 'grey area' that the police were generally reluctant to investigate.

This interviewee said that while their department did not directly provide victims with legal advice, they believed that it might be helpful. On the other hand, they also found that in their experience victims were not often interested in talking to lawyers – what they wanted was mental support, or information on how to change schools, etc. Only those with particularly strong social networks considered legal action, the interviewee said – but the majority just hoped the problem would 'go away' on its own.

F. Interview 6[49]

This interviewee worked for an organization that offers an online counselling service 5 days a week, and is intended to be a place where teens can anonymously receive emotional support via text messaging. The organization also uses online platforms (including Instagram, Facebook, and LiHKG) to connect with teens believed to be at high risk, providing them with information about risks associated with hurtful digital communications (both as a victim and as a perpetrator). While the target group is primarily Cantonese-speaking teens, clients can be as young as 6 and as old as 24. The interviewee reported that (anecdotally) the numbers of clients had continually increased over the last few years. In addition to those who suffered from general online harassment and cyberbullying, this interviewee had also worked with a number of clients who had been blackmailed regarding threatened spread of their intimate images or videos.

In the interviewee's experience, cyberbullying is closely related to 'in-person' bullying in schools, and they believed the latter was becoming an increasing problem

---

[48] Antigone Davis, "Strengthening Our Efforts Against the Spread of Non-Consensual Intimate Images", 2 Dec. 2021, *Meta*, online: https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/.

[49] Interview conducted 13 Jun 2023. Questions asked in English, responses given in Cantonese. The Chinese transcript was then later translated by the research assistant. Both transcripts on file with author.

in Hong Kong due in part to an outflux of experienced educators and social workers from the city. While their organization does offer outreach campaigns in schools, the interviewee believed that generally speaking there was inadequate training for teachers on the specific issue of HDC, and as a result schools were often poor at both identifying and solving HDC-related problems. They reported that student victims are often reluctant to speak with their school social workers because they are worried about secondary harms like developing a reputation in school for complaining. The interviewee reported that this is in part why the online counselling programme is so popular with students, because it is entirely anonymous and is unconnected with anyone in their schools. This interviewee reflected that online counselling was quite different from face-to-face counselling in their experience. It had both advantages (for example, the text-based format meant counsellors had more time to think and reflect before responding) and disadvantages (for example, the text-based format also made it harder for counsellors to read the emotions of their clients and they also sometimes found it hard to gauge the authenticity of the complaints).

In contrast to some other NGOs operating in this space, this organization saw its role primarily as advisory. The interviewee found that many of the users of the service did not seek direct intervention, but instead just wanted a place to receive emotional support or advice. Therefore, they did not have much experience seeking to have material removed from online platforms on behalf of their clients. The interviewee said that while in two or three rare instances they had approached a social media platform and asked for something to be removed, this was unusual and clients understood that this was not their role.[50]

The interviewee explained that while victims of blackmail tended to not want to go to the police, they were advised that sometimes it was necessary if the threats continued to escalate. The interviewee reported that in their experience the police were often unwilling to take on a case unless they were provided with very clear evidence of threats to someone's physical safety. In their view, the police were likely to be receptive to a complaint about cyberbullying or online harassment only if it had such a 'real-world' component. The interviewee was of the view that legal reform leading to a clear offence of 'cyberbullying' would be helpful if it encouraged the police to take it more seriously. They also reflected that smaller NGOs like theirs were somewhat hamstrung by a lack of resources in this area – for instance, they could not hire legal staff to assist clients or even to provide general advice as to their legal rights and options.

G. Interview 7[51]

This was an interview with a representative of a large charitable group in Hong Kong that is involved with a variety of social campaigns, outreach efforts, and provision of medical services. The interviewee worked for a department within this group that had historically focused upon addiction prevention and treatment. In recent years, however, this department had begun to focus more of its resources on 'digital' issues. The focus of the department was primarily on educational campaigns, though in rare cases they offered victims of HDC a limited form of one-on-one counselling. For example, where students were in clear distress about NCDII and associated blackmail, they might be provided with general emotional support alongside advice about what

[50] An explanation as to why these instances were treated differently was not offered.
[51] Interview conducted 19 June 2023. Transcript on file with authors.

could be achieved and how to respond to threats. The department might advise clients when it was necessary to involve the police but would not accompany the students to make reports directly. Similarly, while the department might provide general advice on how to try and obtain removal of images, they did not directly assist victims in that process.

In recent years the department's focus was running an internet addiction campaign and a series of anti-cyberbullying initiatives. The department created educational resources for primary and secondary school students about healthy online behaviours, and also for teachers to help them recognize when cyberbullying might be occurring in their classrooms. The interviewee said that their department was sometimes invited into a school proactively, but more often than not they were engaged only after something had happened and this was – in their view – unfortunate. The interviewee explained that while schools often recruited them in the hope of helping to rebuild relationships between perpetrator and victim, this tended to not solve any underlying issues that led to the incident in the first place.

The interviewee also expressed concern that students were becoming numbed to the issue of HDC, with many just accepting it as an inevitable facet of their lives. In the course of developing the resources for their educational campaigns, the interviewee explained that their department had conducted a survey of students. This survey found that cyberbullying was most common amongst students aged 13-15, and the most common kinds of HDC experienced were the malicious spreading of rumours online and doxing, both of which students described as forms of 'cyberbullying'. Twenty percent of students reported being cyberbullied in the survey, and 30% of those students said that they had experienced some level of suicidal ideation as a result. The survey also revealed that 43% of students who had witnessed cyberbullying of a peer did not intervene (either directly or by reporting the incident to a teacher, parent, or social worker for instance). It also found that the perpetrators of the cyberbullying themselves were often victims previously, showing that there was essentially a cycle of digital violence. The interviewee reported that perpetrators they had spoken with often believed that cyberbullying was a form of enacting justice against a wrongdoer; that it was the right thing to do.

H. Interview 8[52]

The interviewee was a barrister who had provided pro bono services to women who had experienced sexual violence, including HDC. The barrister held the view that victims of image-based / online sexual offences were less open to going to the police than victims of physical abuse and those who were willing faced challenges. The barrister shared examples of their clients experiencing difficulties with pursuing charges, even under the new laws. For instance, one client who was a victim of NCDII but who did not know the identity of the perpetrator went to the police. The client reported that the first officer she spoke to was helpful, and agreed to transfer the case to the Cybercrime Bureau because they might have better resources to identify the perpetrator. But the client claimed that the Cybercrime Bureau then refused to take on her case because she was an adult rather than a minor (although obviously the law draws no distinction here).

[52] Interview conducted 5 September 2023. Transcript on file with authors.

The barrister had only been engaged in relation to criminal matters and so did not have a particular view on civil remedies. However, they indicated that social workers with whom they had spoken reported that the primary interest of their clients who were victims of NCDII was having the relevant material removed from the internet. In terms of legal changes the barrister advocated for practical measures at the police level, such as better training for frontline officers dealing with all kinds of sexual offences. They shared examples of clients who reported victim-blaming or shaming, for instance, where they had taken drugs or voluntarily provided sexual images to another person.

I. Interview 9[53]

The interviewee was a solicitor who in the past had offered pro bono services victims of forms of sexual violence including HDC. A typical example of their work was providing advice to clients on how to purse pursuing conciliation[54] with the perpetrator through the offices of the Equal Opportunities Commission. The solicitor reported that their direct involvement was more likely if the case was particularly strong and the client did not wish to go through the conciliation process, but instead wanted to press for pre-action settlement.

While in the case of 'conventional' sexual harassment in the workplace victims could rely on laws like the Sex Discrimination Ordinance, the solicitor was of the view that the law (in general) was less helpful in the context of newer digital forms of sexual violence. They had no experience with the new disposal orders under the Crimes Ordinance referred to earlier, but were of the view that Hong Kong's data protection laws were of little use. The solicitor described some recent experience they had with assisting clients in obtaining removal of NCDII material. In their view, requests were generally dependent on the goodwill of the webhosts or service providers to voluntarily remove the material upon request, and that was not always the case. The solicitor reported it was generally easy to obtain removal of most explicit non-consensually shared images from larger, mainstream social networks like Facebook or Instagram because they had generic terms of service policies against the posting of nudity or sexual content. However, services like Telegram (that had encrypted channels dedicated to the sharing of NCDII material) were less likely to respond. Where the host or provider was not predisposed to assist, the solicitor described their legal options as limited and challenging – especially if they were based outside Hong Kong. As explained above, courts required an image make reference to a specific individual for it to be classed as personal data, and rarely would someone upload an intimate image with the persons' real name attached. Sometimes the images also just showed parts of the body with no clear face, and so no identity inference could be drawn.

The solicitor doubted whether the courts would respond positively to an action in tort. In the case of NCDII, clients typically first wanted the images taken down, second some kind of apology from the perpetrator, and only then might they start thinking about other kinds of remedies. Monetary compensation was rarely a client's focus, and so the solicitor said in the past they had considered obtaining an injunction as a possible method of preventing the further spread of intimate images. However, this

---

[53] Interview conducted 4 Oct 2023. Transcript on file with authors.

[54] See "What is conciliation?", *Equal Opportunities Commission of Hong Kong*, online: https://www.eoc.org.hk/en/enquiries-and-complaints/what-is-conciliation.

proved difficult as a court requires specific details about subject-matter before an injunction will be granted regarding it. In practice the solicitor's clients were often unsure about the number of images that existed or what precisely was contained in each one. It was also challenging to identify the respondent towards whom the injunction should be directed – frequently the images were hosted on servers outside of Hong Kong, and whoever uploaded them did so anonymously. These challenges meant the solicitor would generally advise their clients that this kind of legal approach was not likely be to an effective solution, especially if the image had already spread beyond the larger mainstream services.

The solicitor considered that having some kind of governmental or semi-governmental organization able to assist in matters of takedown, apology, and compensation might be helpful. They also suggested that a clearer division of labour amongst existing government bodies (eg the police, Equal Opportunities Commission, PCPD) about who was responsible for what, or at least it being clear as to which body a victim could approach for assistance, who they could file a complaint with, etc, would be a step in the right direction.

IV. ANALYSIS

The above interviews reveal several patterns. First, they suggest that incidents of HDC are on the rise in Hong Kong. While most of the NGOs with which we spoke did not maintain detailed statistics, they were consistent in reporting anecdotal shifts in the kinds of cases that were coming before them. These reports are consistent with the various surveys referred to in the opening section of this paper. Several representatives noted a particular increase noted since the COVID-19 pandemic. While we cannot definitively prove it here, it is at least plausible that the increase may be connected to the long school closures and social distancing rules imposed on students in Hong Kong during the pandemic, which led students to spend increasing amounts of time online.[55]

Second, one can note some commonalities regarding the characteristics of victims who approach the NGOs. Again though they generally do not maintain detailed statistics, the representatives with whom speak consistently reported that the majority of their clients who have experienced HDC are under 25. Of course, there are two important caveats to this – first (and again) this is anecdotal reporting. Second, the age profile of the clients is in a sense self-selecting depending on the focus of the NGO in question. That is, if an NGO sets out to provide services to youth then obviously most of its clients will be young. However, we came across no reports or surveys that suggests there are large numbers of middle-aged or elderly people who are victims of HDC. That does not mean that no such victims exist of course, but simply that surveys, reports, and discussions with stakeholder groups all point towards the problem of HDC in Hong Kong being youth-oriented. We also find that gender does not seem to be a determinative factor, with multiple groups reporting dealing with both male and female victims and perpetrators. However, on balance it does seem that victims are more likely

[55] See e.g., Min Lan, Qianqian Pan, Cheng Yong Tan, and Nancy Wai Ying Law, "Understanding protective and risk factors affecting adolescents' well-being during the COVID-19 pandemic", (2022) 7(32) NPJ SCIENCE OF LEARNING https://doi.org/10.1038/s41539-022-00149-4 and Albert Lee, Vera MW Keung, Vincent TC Lau, Calvin KM Cheung, and Amelia SC Lo, "Impact of COVID-19 on Life of Students: Case Study in Hong Kong" (2021) 18(9) INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH 10483 https://doi.org/10.3390/ijerph181910483.

to be female. But again, the same caveats apply regarding the lack of clear statistical record-keeping.

A third pattern revealed in the interviews regards the *form* of HDC – while this is dependent on both accurate reporting by a victim and recording by an NGO, it appears that cyberbullying and NCDII are the two most common forms of HDC seen by local stakeholder NGOs. This may be because that the phrase 'cyberbullying' can function as a kind of blanket descriptor that can include harassing messaging as well as doxing and blackmail. While cyberbullying understood in that broad sense could also include NCDII (especially if it leads to blackmail), it nonetheless appears that the interviewees tended to treat the latter as a separate (if related) problem. For that reason that breakdown is adopted in the section that follows when considering internal patterns within those two categories revealed by the interviews.

The interviews revealed several patterns within the broader 'cyberbullying' category. First, cyberbullying is often connected to 'real world' bullying in schools rather than something that is an isolated phenomenon. Second, it is often part of a cycle of digital violence – victims frequently go on to become perpetrators themselves. Third, a number of the interviewees who worked for stakeholders that focused on schools suggested that schools are too often taking a reactive rather than proactive approach. They also suggested that training on cyberbullying should be provided to not only students, but also to teachers and staff. Fourth, several stakeholders suggested that the rise of end-to-end encryption in messaging apps poses a new problem that reinforces the need to prevent HDC from occurring in the first place as these apps can make it very hard to detect if students decline to report the problem. Fifth, the primary concern of victims is that the cyberbullying end and the maintenance of their social status rather than legal prosecution or punishment of the perpetrators. Finally, multiple stakeholders reported that the police were relatively unresponsive to complaints about cyberbullying unless it was accompanied by threats to physical safety or demands for money. In other words, the police often appeared to perceive their role as primarily confined to the 'real world', at the expense of the digital.

The interviews also reveal clear patterns related to NCDII. First, the interviews suggest that NCDII primarily occurs in three situations – in the context of sex work where a client secretly records a sex worker without consent; the breakdown of an existing intimate relationship where one party threatens to spread images that were initially consensually created; or where someone has been convinced or coerced into sending intimate images to someone they only known online. Second, while sometimes NCDII is done purely for 'revenge', it appears from the interviews that more often it is done for the purposes of blackmail. That blackmail can take various forms, including demands for money or sexual favours. The threat preys on the victim's fear of embarrassment, shame, or personal or professional loss because of what those images would reveal about the victim's sexual orientation, preferences, or behaviours. Third, multiple stakeholders considered the new criminal provisions specifically targeting NCDII to have limited value. Several interviewees reported that victims did not trust that their complaint would be handled empathetically by the police – they feared being victim-blamed or shamed. It also appeared that the police were more responsive to complaints when they were made on behalf the victims or when the victims were accompanied by the interviewees to make these complaints. But even then, it appears that the reluctance was also driven in part by fear on the part of victims of a long drawn

out legal procedure would make the problem worse, by increasing the notoriety of the images or lengthening the ordeal – what they really wanted was quick and easy removal of the material from the internet. As with cyberbullying above, what victims appear to want most is elimination of something that appears to threaten their social standing or relationships. Punishment of the perpetrator is a secondary concern.

On the matter of attempting to have offensive material removed from the internet, more patterns appear in the interviews. Notably, so-called 'take-down' was only treated as an option in the context of sexually explicit images or videos – none of the interviewees reported any attempts to have offensive text-based HDC (such as harassing or defamatory words or threats) removed from websites. In the context of NCDII, interviewees consistently reported that mainstream multinational social media services (with the exception of X – this was noted on several occasions) tended to respond relatively positively and quickly to removal requests relating to NCDII material. Those services tended to already have 'safety' teams in place and clear policies regarding not just NCDII but sexually explicit material in general. Some of the NGOs had developed productive working relationships with these teams. In contrast, niche websites devoted to pornography or more local websites were more likely to ignore requests, and lacked clear procedures about how to go about making a request. A number of interviewees also stated that the use of encrypted messaging apps created another challenge that was as yet unresolved.

CONCLUSION

This paper has described the first stage of an important project aimed at the appropriate regulation and control of hurtful digital communications in Hong Kong. Interviews with representatives of nine local stakeholders suggest that the problem of HDC is on the rise in Hong Kong and the introduction of some related criminal penalties around NCDII and doxing have not solved and do not appear likely to solve this problem. The next stage of the project will engage in a comparative analysis of tools adopted in other jurisdictions including more expansive private law remedies and broader regulatory regimes that seek to create expedited administrative procedures for victims of HDC. Of the former, consideration will be given to the development elsewhere of judicially-developed and statutory torts related to invasions of privacy and harassment. Of the latter, examples might include the powers and approaches of Australia's E-Safety Commissioner[56] or the online Civil Resolution Tribunal[57] in the Canadian province of British Columbia, both of which attempt to offer easier mechanisms for individuals seeking removal of certain categories of harmful material online. The UK's Online Safety Act may also prove instructive, as it mandates that all "user to user" online services (such as social media) have processes in place that allow for the quick removal of all kinds of illegal content.[58] Likewise, Canada's proposed Online Harms Act envisions that all platforms will have to systems in place to remove

---

[56] *E-Safety Commissioner, Government of Australia*, online: https://www.esafety.gov.au.

[57] "Intimate Images", *Civil Resolution Tribunal, Government of British Columbia*, online: https://civilresolutionbc.ca/solution-explorer/intimate-images/.

[58] Online Safety Act, 2023 c. 50, s. 10, online: https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted. The precise mechanisms of how this will work in practice are currently being developed by the regulatory authority, *Ofcom*. See "Consultation: Protecting People from Online Harms", Ofcom, 9 Nov. 2023, online: https://www.ofcom.org.uk/consultations-and-statements/category-1/protecting-people-from-illegal-content-online.

certain categories of illegal material within 24 hours after being notified.[59] These examples are by no means exhaustive of the options that exist, but point the direction of research into alternative mechanisms of control.

To reiterate, the first stage of the project as described in this paper has been to begin to better understand patterns of HDC as they occur in Hong Kong. This was necessary a first step because we do not propose that a legislative solution applied elsewhere can simply be imported directly. Any solution must take into account the local context, including not only the specific legal and political environment but also a proper understanding of the nature of HDC practices as they occur locally. The third and final stage of this project will leverage the comparative research gleaned from the second stage in order to inform recommendations for regulatory reform in Hong Kong that take into account this context. In other words, to what extent can overseas experience help develop solutions that effectively redress the specific harms Hong Kongers face as a result of hurtful digital communications? There is unlikely to be a 'magic bullet' solution, but it is clear that the present situation is untenable.

[59] Bill C-63: An Act to enact the Online Harms Act, to amend the Criminal Code, the Canadian Human Rights Act and An Act respecting the mandatory reporting of Internet child pornography by persons who provide an Internet service and to make consequential and related amendments to other Acts, s. 67, online: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading.

REFORMING CORPORATE COMPLIANCE SYSTEMS IN CHINA UNDER THE NEW COMPANY LAW: LESSONS FROM THE UNITED STATES AND JAPAN

Zongqi Wu*

Abstract: This article investigates the construction of an ideal compliance system under China's new Company Law through a comparative legal analysis. It begins by identifying the existing challenges in China's compliance framework, particularly the historical dominance of shareholder primacy and the uneven distribution of responsibilities under the previous Company Law. These issues have hindered the development of an effective compliance system, necessitating a reevaluation of corporate governance principles in China. The comparative analysis examines the compliance systems in the United States and Japan, focusing on the frameworks and key elements such as fiduciary duties and the application of the business judgment rule. The study highlights significant differences in how these jurisdictions handle compliance, providing insights into potential improvements for China's system. This section underscores the importance of adapting international best practices to local contexts to enhance the effectiveness of corporate compliance in China. Finally, the article proposes a dual-level compliance concept tailored to China's unique legal and corporate environment, integrating both private and public interests. It outlines specific compliance duties for directors, supervisors, and other compliance personnel, emphasizing the need for robust information transmission, whistleblower protections, and temporary management systems. The conclusion calls for a comprehensive and flexible compliance framework that not only ensures corporate profitability but also promotes broader societal welfare, aligning with the evolving goals of corporate governance in China.

Keywords: Compliance System; Corporate Governance; Fiduciary Duties; Business Judgment Rule; New Company Law

* The Chinese University of Hong Kong, China.

Table of Contents

Introduction 149

I. Overview of Corporate Compliance Systems in China 150

A. Definition of the Compliance Concept in China 151

B. China's Compliance Challenges 151

1. The Dominance of Shareholder Primacy in China 152

2. Lack of Obligations for Key Roles in the Old Company Law 153

II. Innovation in Corporate Compliance Systems Under China's New Company Law 155

A. Establishment of Employee Representatives 155

B. Innovation in the Duties and Responsibilities System for Directors 156

1. Strict Liability of Directors to Third Parties 156

2. China's Version of the Shadow Director System 157

III. Comparative Analysis of Corporate Compliance Systems 157

A. The Compliance System Framework Under U.S. Law 157

1. Overview and Evolution of Compliance Systems 158

2. Key Milestones in the Evolution of Compliance Systems .. 159

3. Compliance Obligations of Directors 161

4. Application of the Business Judgment Rule 163

B. The Compliance System Framework Under Japanese Law 164

1. Legal Basis 164

2. Basic Elements of the Internal Control System 165

3. The Duties of Directors 168

4. Application of the Business Judgment Rule 169

C. Horizontal Comparison 170

IV. Establishing an Ideal Compliance System Model 171

A. Duality of Compliance Concepts 172

1. Consideration of Private Interests 172

2. Public Interest 174

B. Typification of Corporate Compliance Rules 174

1. Directors' Compliance Duties 175

2. Supervisors' Compliance Obligations 178

3. Obliations of Other Compliance Personnel 178

C. Composition of the Compliance System 180

1. Connective Elements of the Compliance System 180

2. Functional Elements of the Compliance System 182

Conclusion 184

INTRODUCTION

Non-compliance with laws and regulations can lead to significant reputational harm for organizations, investments in compliance have kept up with the evolving regulations and enforcement measures.[1] Hence, the enactment of the new Company Law in China marks a significant shift in the nation's corporate governance landscape, emphasizing the importance of robust compliance systems. [2] Historically, China's corporate governance has been dominated by the principle of shareholder primacy, which prioritizes shareholder interests often at the expense of broader stakeholder considerations.[3] This focus has led to challenges in enforcing compliance, with a lack of clear frameworks and responsibilities resulting in inconsistencies and inefficiencies. The new Company Law aims to address these issues by introducing comprehensive reforms designed to balance shareholder interests with broader corporate responsibilities.[4]

In China, the concept of compliance has traditionally been fragmented and inconsistently applied.[5] The previous Company Law provided limited guidance on the distribution of compliance responsibilities, leading to ambiguities and varied interpretations across different regions and industries.[6] This inconsistency has been compounded by the lack of a standardized judicial approach to adjudicating cases of director diligence and fiduciary duty, making it difficult to enforce compliance uniformly.[7] The new Company Law seeks to rectify these shortcomings by establishing clear guidelines and responsibilities for corporate directors, supervisors, and other key personnel.

To contextualize these reforms, it is essential to examine the compliance frameworks of other jurisdictions, particularly the United States and Japan. The United States offers a well-established model of corporate compliance, with fiduciary duties

---

[1] Harald Haelterman, Breaking Silos of Legal and Regulatory Risks to Outperform Traditional Compliance Approaches, 28 Eur. J. Crim. Pol'y & Res. 19, 20 (2022).

[2] Arendse Huld, China's Revised Company Law in Effect from July 1, 2024: Key Details Here, China Briefing (June 8, 2024), https://www.china-briefing.com/news/china-company-law-amendment-july-1-2024/.

[3] Robert J. Rhee, A Legal Theory of Shareholder Primacy, 102 Minn. L. Rev. 1951, 1951 (2017).

[4] Hawksford, Explaining the recent amendments to China's Company Law, Hawksford (March 12, 2024), https://www.hawksford.com/insights-and-guides/explain-amendments-to-china-company-law (Fiduciary and diligence obligations for senior executives).

[5] There is no clear and uniform definition of compliance within China's academic circle; different scholars have different opinions, resulting in various interpretations of the concept and respective approaches. The lack of a unified opinion has made it hard to have a standardized compliance framework for theoretical research and its practice in different business sectors. See, e.g., Zhao Wanyi (赵万一), Hegui Zhidu de Gongsi Fa Sheji Jiqi Shixian Lujing (合规制度的公司法设计及其实现路径) [The Design and Implementation Path of Corporate Compliance Systems in Company Law], 2 Zhongguo Faxue (中国法学) [China L.] 69, 71 (2020) (China).

[6] Deng Feng (邓峰), Gongsi Hegui de Yuanliu Ji Zhongguo de Zhidu Juxian (公司合规的源流及中国的制度局限) [The Origins of Corporate Compliance and Institutional Limitations in China], 1 Bijiao Fa Yanjiu (比较法研究) [Comp. L. Res.] 34, 42 (2020) (China).

[7] One of the most important things will be a unified legal concept in use and interpretation across several sectors in China. It is an increment to coherence and clarity on the application and interpretation of the said laws. A cohesive legal framework could achieve better legal predictability, stable business conditions, and greater potential for fair and equal treatment before the law. Id.

and the business judgment rule playing crucial roles in ensuring accountability.[8] This model encourages directors to make bold and innovative decisions while being protected from legal repercussions, provided they act in good faith and with due diligence. Japan, on the other hand, integrates compliance within its broader internal control systems, focusing on the duty of care and loyalty of directors and emphasizing a holistic approach that includes stakeholder interests and robust internal controls.[9]

Drawing insights from these international experiences, this article proposes a dual-level compliance concept tailored to China's unique legal and corporate environment.[10] This concept integrates both private and public interests, ensuring that compliance duties are clearly defined and enforced across all levels of corporate governance.[11] The proposed model emphasizes the importance of robust information transmission systems, whistleblower system, monitoring system and business screening system to enhance the effectiveness and resilience of the compliance framework.[12]

The new Company Law seeks to redefine the roles and responsibilities of corporate directors, supervisors, and other key personnel, ensuring that compliance obligations are clearly delineated and systematically enforced.[13] This redefinition is essential for addressing the historical dominance of shareholder primacy and fostering a more balanced and effective compliance system. By incorporating international best practices and adapting them to the local context, China can establish a compliance system that not only meets legal standards but also promotes sustainable and ethical corporate governance.

In conclusion, the construction of an ideal compliance system under China's new Company Law requires a comprehensive and flexible framework that balances corporate profitability with societal welfare. This article aims to contribute to the ongoing discourse on corporate compliance in China, providing a roadmap for future reforms and advancements in corporate governance.

I. OVERVIEW OF CORPORATE COMPLIANCE SYSTEMS IN CHINA

The concept of a compliance system and its practical application have garnered significant attention in Chinese academic circles.[14] However, given China's previous legislative environment, effectively integrating a compliance system has proven challenging. This section will discuss China's understanding of the compliance concept and the reasons behind the difficulties in building a compliance framework in China.

---

[8] Id. at 35-38.
[9] See infra section II.B.3, pp. 28-29.
[10] See infra section III.A, pp. 33-36.
[11] Id.
[12] See infra section III.C.1, pp. 45-46.
[13] Jonathan Bench, Understanding the China Company Law Amendments that Matter to Foreign Businesses, Harris Sliwoski (May 15, 2024), https://harris-sliwoski.com/chinalawblog/understanding-the-china-company-law-amendments-that-matter-to-foreign-businesses/.
[14] See Zhao Wanyi supra note, at 69.

A. Definition of the Compliance Concept in China

Currently, the interpretation of "compliance" is difficult to unify in Chinese, primarily due to the diverse forms and varying applications of the legal departments it involves. [15]In an English context, "compliance" means "to comply with or are asked to do," initially referring to adherence to effective regulations or performing required actions. [16] In China, however, the academic community has many definitions of compliance, with a more authoritative explanation found in Article 2 of the "Compliance Management Measures for Central Enterprises." This defines compliance as the behavior of enterprise operations and management, as well as the performance of duties by employees, in accordance with national laws and regulations, regulatory requirements, industry standards, international treaties, rules, and company charters and related regulations.[17]

Regardless of the definition adopted, it is impossible to encapsulate the rich connotations of the concept of a compliance system. I think, understanding the concept of compliance in China should be approached from two dimensions. First, as an external control role, it requires that both internal and external behaviors of a company comply with laws, regulations, and mandatory norms while simultaneously meeting commercial practices, internal company behavior consistency, and company bylaws. This represents the external requirements and expectations for company operations. Second, as an internal operational self-restraint, it aims for the company to adapt to changes in the internal and external environment to achieve stable self-control and sound management.[18]

In summary, the concept of compliance in China is neither solely the control by external legislation nor merely confined to the internal autonomy of the enterprise.[19] Instead, it should be an organic combination of both, aimed at the effective management of the company and the efficient governance of society.[20] This point reveals that in China, the concept of corporate compliance also has obvious ESG characteristics.

B. China's Compliance Challenges

China's compliance system has faced many difficulties in its construction. The main reasons are the prevalence of shareholder primacy in China and the unequal distribution of responsibilities under the old Company Law.[21]

---

[15] See Id. at 71.
[16] Id.
[17] Zhongyang Qiye Hegui Guanli Banfa (中央企业合规管理办法) [Compliance Management Measures for Central Enterprises] (promulgated by the State-owned Assets Supervision and Administration Commission, Aug. 23, 2022, effective Oct. 1, 2022), State-owned Assets Supervision and Administration Commission Order No. 42, available at https://www.gov.cn/zhengce/zhengceku/2022-09/19/content_5710633.htm (China).
[18] See Zhao Wanyi supra note 7.
[19] Compliance, in a sense, also aligns with the requirements of ESG principles. Id.
[20] Id.
[21] Deng Feng supra note 6, at 43.

1. The Dominance of Shareholder Primacy in China

Shareholder primacy, also known as shareholder-centric, posits that shareholders occupy the most central position in the entire corporate organizational structure.[22] They are considered the owners of the company and can employ professional managers to operate the company for the benefit of the shareholders.[23] In a company, shareholders entrust professional managers to manage the company on their behalf, with the implicit default rule being that professional managers should utilize their management skills to help shareholders achieve their profits.[24] Consequently, in the course of management, professional managers have no reason to constrain themselves for any purposes other than those of the shareholders. Furthermore, the ultimate power of the company is exercised by the shareholders. Even though, in most cases, shareholders may not bypass the governance expertise provided by professional managers, they still retain the final decision-making authority.

In China, most companies strictly adhere to the principle of shareholder primacy.[25] In routine governance, the voting rights represented by shareholders and the management rights represented by the board of directors often appear inseparable, a classic phenomenon.[26] The organizational level of companies is relatively low, with almost all company powers concentrated in the general meeting of shareholders.[27] The board of directors is practically regarded as an extension of the shareholders' meeting, lacking compliance management. The development of China's corporate system is relatively recent, evolving from Sino-foreign joint ventures.[28] In such joint ventures, to ensure the stability of company development and with the frequent inclusion of state-owned capital, the equity structure is highly concentrated.[29] This equity-centric ideology has continued to this day, whether in limited liability companies with a prominent "personality" attribute or joint-stock companies with a significant "capital" attribute. The power of the general meeting of shareholders (and major shareholders) has always been central, directly overlooking the representative system characteristics that should be met in corporate governance.[30]

The excessive concentration of power in the general meeting of shareholders or a particular shareholder has blurred the essential feature of the company as an independent entity.[31] Some Chinese perspectives have even seriously deviated from the essence of the company, viewing the company as the property of the shareholders,

---

[22] D. Gordon Smith, The Shareholder Primacy Norm, 23 J. Corp. L. 277, 282 (1997).
[23] Id.
[24] Id.
[25] Liu Kaixiang (刘凯湘) & Liu Jing (刘晶), Woguo Gudonghui Zhongxin Zhuyi de Lishi Chengyin (我国股东会中心主义的历史成因) [The Historical Causes of Shareholder Centralism in China], 6 Faxue Luntan (法学论坛) 51, 52 (2021) (China).
[26] Id.
[27] Deng Feng supra note 13.
[28] Id.
[29] Id.
[30] Id.
[31] Excessive concentration of power within the shareholders' meeting, whether directly or indirectly, results in the loss of independence of a company. All such imbalances innovate against the basic principles of good corporate governance—checks and balances—which could consequently result in decisions favoring just a few at the expense of other broader stakeholder interests. Id.

who can freely dispose of the company's assets and personnel. This leads to the disregard of existing election procedures. Shareholders, relying on the "principal-agent" system, can effectively replace directors and senior management at any time.[32] In state-owned enterprises, the majority implement a party committee system, where the appointment and removal of senior management are often overridden by party regulations.[33] Government agencies, as shareholders, seem to be exempt from or find it challenging to comply with the basic limitations of internal company charters and procedures.[34] This situation actually contradicts the provisions of company law, which stipulate that the powers of the board of directors are derived from the general meeting of shareholders through elections.[35]

Although under the principle of shareholder primacy, as discussed above, the positions of board members, supervisors, and senior management within a company are not stable, in the process of accountability, these agents in the principal-agent relationship often passively bear the responsibility for executing the will of the principals. In fact, from my perspective, the external form of responsibility is generally borne by the company making the external declarations, while internal responsibility should be allocated according to different divisions of labor. However, under this theoretical model, the internal declarations of the company are predominantly influenced by the shareholders' will. It is challenging to separate the will of senior management from that of the shareholders. In some instances, management personnel may have no involvement of their own will in certain decisions. According to general principles of responsibility allocation, the shareholders behind these decisions should logically be held accountable. However, in practice, it is very rare for such accountability to be pursued in China.[36]

2. Lack of Obligations for Key Roles in the Old Company Law

In the corporate governance process in China, the primary duty of directors is fiduciary duty, which comprises two categories.[37] The first is the duty of loyalty, meaning that in the course of corporate governance, directors must prioritize the interests of the company and must not place personal interests above those of the company.[38] The second is the duty of care or duty of diligence, which requires directors to act prudently in the course of corporate governance to prevent actions that may lead to losses for the company.[39] These two duties form the basis of directors' responsibility

---

[32] Id.
[33] Liu Kaixiang & Liu Jing supra note 17, at 54.
[34] Id.
[35] Liu Kaixiang & Liu Jing supra note 17, at 52.
[36] In China, companies controlled by shareholders effectively do not have independent status, making actual claims against the shareholders hard to pursue. This scenario engenders a shock to principles of corporate autonomy and accountability, giving rise to possible cases of conflicts of interest and corporate governance. Deng Feng supra note 13, at 44.
[37] Fu Qiong (傅穹), Gongsi Liyi Fanshi Xia de Dongshi Yiyi Gaige (公司利益范式下的董事义务改革) [Reforming Directors' Duties Under the Paradigm of Corporate Interests], 6 Zhongguo Faxue (中国法学) [China L.] 197, 204 (2022) (China).
[38] Id.
[39] These two duties regulate the behavior pattern of directors from two dimensions. They set clear guidelines to ensure the best conduct of the company's interest with a touches of ethical standards. Id.

to the company in their management roles.[40]

However, despite this seemingly comprehensive duty system, there are significant issues. The construction of China's Company Law differs from that of common law countries such as the UK and the US, adopting legislative principles more aligned with French law. French law, in its requirements for company directors, places greater emphasis on the fulfillment of the duty of loyalty during corporate governance, while the demands for the duty of diligence are relatively lower. Consequently, although the 2005 Company Law of China stipulates the duties of directors, it does not provide detailed explanations on how directors should fulfill their appropriate duties or what constitutes the fulfillment of such duties.

China previously lacked the foundational basis for establishing a compliance system. This article argues that the construction of a compliance system may include not only the establishment of specialized compliance organizations and the clarification of compliance norms but, regardless of how the final system is constructed, the most critical aspect is rationally establishing a complete set of obligations. I believe that the clarification of these obligations is the cornerstone for constructing the entire compliance system, with two main adjustment directions:

First, the system should include how the compliance department ultimately transmits or executes the necessary compliance information. During the operation of the compliance department, any information potentially harmful to the company's interests must be promptly communicated to the company's directors, supervisors, and other senior management personnel.

Second, the construction of the compliance obligations also plays a crucial role in the allocation of internal responsibilities within the company and the assumption of external responsibilities. Clear obligations are essential for the substantive allocation of responsibilities. The aforementioned obligation system had significant loopholes before the introduction of the new Company Law, preventing China from constructing an effective compliance system.

Simultaneously, the lack of uniformity in judicial decisions across different regions in China also complicates the construction of a compliance system.[41] The current judicial approach in China regarding whether directors have violated the duty of diligence generally follows these steps: first, establishing a causal relationship between the objective losses specified by the company and the directors' actions, excluding situations arising from normal business risks or force majeure; second, determining whether the directors' actions violated internal company bylaws or laws and regulations; and finally, examining whether the directors' actions were prudent and reasonable.

---

[40] Id.

[41] For example, the consequence of local protectionism can be different court rulings on the very same case. Such inconsistency undermines equal justice under law and may eventually destroy public confidence in the rule of law. Yuan Meng, The Limits of Judicial Reforms: How and Why China Failed to Centralize Its Court System, 255 China Q. 753, 763 (2023).

Although this "three-step" judgment approach appears to be progressive, there are two main issues faced by judges in the adjudication process:

First, regional differences. For corporate disputes in different regions, particularly in economically more developed and mature market areas,[42] the business models are more complex, and naturally, the requirements for directors' duty of diligence are higher. Consequently, the standards for adjudicating cases vary according to the level of economic development and market environment of different regions. This results in differing standards for examining directors' duties and the duty of diligence, leading to regionally stable but inconsistent judicial outcomes.

Second, there are few cases where courts adjudicate based on the legal principle of the duty of diligence. This lack of a unified rule for determining whether the duty of diligence has been violated leads to inconsistencies, with some judges even conflating the duty of loyalty with the duty of diligence. This severely impacts the stability of judicial decisions.

II. INNOVATION IN CORPORATE COMPLIANCE SYSTEMS UNDER CHINA'S NEW COMPANY LAW

China currently lacks requirements for establishing compliance systems, and, as previously discussed, this presents significant deficiencies. In the context of the new Company Law, China aims to lay the groundwork for a more comprehensive compliance system and further pursue ESG goals by making improvements related to employee rights and the duties and responsibilities of directors.[43]

A. Establishment of Employee Representatives

The new Company Law in China has removed the upper limit on the number of directors and has further improved the system for establishing employee representatives. Previously, only certain wholly state-owned enterprises and some limited liability companies were required to have employee representatives. Under the new Company Law, this requirement has been expanded to include all limited liability companies with more than 300 employees. Moreover, it emphasizes the allowance for the inclusion of

---

[42] Wenyan Zhu (朱文雁), Cong Sifa Xianzhuang Kan Panli Zai Woguo Sifa Guocheng Zhong de Jiazhi (从司法现状看判例在我国司法过程中的价值) [The Value of Case Law in China's Judicial Process from the Perspective of Current Judicial Status], 21 Faxue Luntan (法学论坛) 120, 123 (2006) (China).

[43] Lei Zhao (赵磊), Gongsi Fa Shang Xinyi Yiwu de Tixi Goucheng——Jian Ping Xin "Gongsi Fa" Xiangguan Guiding (公司法上信义义务的体系构成——兼评新《公司法》相关规定) [The Structure of Fiduciary Duties in Company Law—Comments on the Relevant Provisions of the New Company Law], 3 Caijing Faxue (财经法学) 67, 67 (2024) (China).

employee representatives in boards of directors consisting of three or more members.[44] This enhancement strengthens the interests of employees and other stakeholders, significantly improving the compliance foundation at the employee level.

B. Innovation in the Duties and Responsibilities System for Directors

As previously described, the main obstacle to building a compliance system in China has been the unclear duties and responsibilities of directors and other personnel. Therefore, in the new Company Law, China addresses this by establishing effective duties to compensate for these shortcomings.[45]

1. Strict Liability of Directors to Third Parties

China's new Company Law legislatively imposes strict duties and responsibilities on directors and other senior management personnel for harm caused to third parties.[46] Previously, the Company Law only stipulated compensation rules for damages caused to the company by directors and senior management personnel in the execution of their duties, without clearly defining their liability to third parties. The new Company Law explicitly outlines the directors' responsibility to third parties, thereby strengthening the duties and responsibilities of directors, which are critical to the integrity of the compliance system.[47]

However, I hold the view that this new regulation may be somewhat unreasonable as it could excessively expand the directors' liabilities. In the business world, no decision can satisfy everyone, implying the existence of both beneficiaries and those adversely affected. Therefore, directors should not be held liable for every third party's losses resulting from their business decisions. Even if subjective elements are limited, this remains ineffective because the underlying logic of any business decision involves intentional or gross negligence, which falls outside the reasonable scope of liability. Thus, while this regulation appears to strictly protect third-party

---

[44] By having employee directors, companies can definitely ensure the safeguarding of the rights of workers to some extent. In other words, through this, workers obtain representation in decision-making, and the company's governing structure becomes fairly comprehensive and whole. This method reduces the possibility of concerns arising from the employee end and increases job satisfaction by developing a sense of ownership and dedication among the workers, thereby contributing immensely to overall company stability and success. Zhao Xudong (赵旭东), Zhou Linbin (周林彬), Liu Kaixiang (刘凯湘), Zhao Wanyi (赵万一), Zhou Yousu (周友苏) & Li Jianwei (李建伟), Xin "Gongsi Fa" Ruogan Zhongyao Wenti Jiedu (Bitan) (新《公司法》若干重要问题解读（笔谈）) [Interpretation of Several Important Issues in the New Company Law (Symposium)], 2 Shanghai Zhengfa Xueyuan Xuebao (Fazhi Luncong) (上海政法学院学报(法治论丛)) 1, 25 (2024) (China).

[45] Zhao Lei supra note 35, at 71.

[46] The responsibility of directors towards third parties has been reconstructed from the perspective of tort theory. This approach redefines the scope and nature of directors' duties, emphasizing their accountability for actions that cause harm to external parties. Tang Xin (汤欣) & Li Zhuozhuo (李卓卓), Dongshi Dui Disanren Zeren de Lilun Jichu Yu Guifan Gouzao (董事对第三人责任的理论基础与规范构造) [Theoretical Basis and Normative Structure of Directors' Liability to Third Parties], 3 Falü Shiyong (法律适用) 75, 76 (2024) (China).

[47] Id.

interests, it may, in practice, lead to an unreasonable expansion of directors' liabilities.

2. China's Version of the Shadow Director System

The new Company Law in China clarifies the duties and responsibilities of controlling shareholders and actual controllers, emphasizing their obligations towards the company's social responsibilities. In common law systems, the shadow director doctrine is used to regulate the unlawful behaviors of controlling shareholders and actual controllers by subjecting them to directors' duties and responsibilities.[48] China's new Company Law adopts a similar theory but extends the civil law concept of joint tort liability, which holds those who instigate or assist in unlawful acts jointly liable.[49] This extension aims to hold controlling persons jointly liable with directors and senior management for unlawful acts.[50] This regulation helps prevent the abuse of control rights by protecting the interests of the company and its minority shareholders from improper conduct by controlling persons.[51]

III. COMPARATIVE ANALYSIS OF CORPORATE COMPLIANCE SYSTEMS

This section provides a detailed comparative analysis of corporate compliance systems in the United States and Japan. It highlights the key components and structures of these systems, focusing on fiduciary duties, business judgment rules, and internal control mechanisms. The U.S. compliance framework emphasizes fiduciary duties and the business judgment rule, encouraging bold decision-making while ensuring accountability. Japan's approach integrates compliance within broader internal control systems, emphasizing the duty of care and loyalty of directors and considering stakeholder interests. By examining these international models, the section draws insights to propose a dual-level compliance concept tailored to China's unique legal and corporate environment, ensuring effective governance and balancing private and public interests.

A. The Compliance System Framework Under U.S. Law

The U.S. compliance system involves professionals such as corporate lawyers and auditors to oversee corporate operations, emphasizing self-regulation to prevent violations and unethical practices.[52] Developed over forty years, this system addresses

---

[48] Wei Zhang (张巍), Dongshi Dui Gudong You Wu Xinyi Yiwu——Xin Gongsi Fa Di 191 Tiao, Di 192 Tiao Zhi Fali Yu Xianshi (董事对股东有无信义义务——新公司法第一百九十一条、第一百九十二条之法理与现实) [Do Directors Have Fiduciary Duties to Shareholders? The Theory and Reality of Articles 191 and 192 of the Ne-w Company Law], 4 Renmin Sifa (Yingyong) (人民司法（应用)) 106, 109 (2024) (China).

[49] Tang Xin & Li Zhuozhuo supra note, at 38.

[50] Zheng Yu (郑彧), Shizhi Dongshi de Falü Guizhi: Yin He, Wei He Yu Ru He (实质董事的法律规制：因何、为何与如何) [Legal Regulation of De Facto Directors: Reasons, Purposes, and Methods], 3 Caijing Faxue (财经法学) 81, 95 (2024) (China).

[51] Id.

[52] Arthur Andersen LLP v. United States, No. 04-368, 2005 U.S. LEXIS 4348, at *1008-1019 (U.S. May 31, 2005).

the inadequacies of corporate self-regulation highlighted by historical financial crises. Key components include connecting structures, top-level awareness, and a compliance culture, supported by specialized employees. These elements ensure robust compliance through mandatory disclosures and fostering a compliance-oriented corporate environment, crucial for effective governance.

1. Overview and Evolution of Compliance Systems

In the United States, the compliance system generally involves professional personnel such as corporate lawyers, auditors, and ethics officers to help supervise the basic operations of companies. In capitalist countries, financial crises are sometimes seen as inevitable cycles brought about by their economic systems. For the U.S. federal government, the numerous corporate illegal activities and crimes during these cyclical crises are closely related to inadequate self-regulation by companies. [53]Since the 1970s, scholars have suggested constructing internal management structures or decision-analysis institutions to help companies avoid violations, unethical business practices, and illegal actions. [54]Subsequently, scholars began advocating for the establishment of compliance management systems to achieve corporate self-regulation, including enforced self-regulation and management-based regulation.[55]

However, research in the early 21st century indicated that due to the inherent biases in the implementation of corporate management measures, companies were often unable to make necessary changes to achieve external policy goals,[56] resulting in compliance systems failing to reach their ideal state.[57] In some cases, corporate managers might spend a lot in forming a compliance system with high cost, still resulting in hard to deter wrongdoing,[58] who are more inclined to accept the high risks of non-compliance to achieve higher expected benefits, rather than adopting compliance management as intended by regulatory bodies. Even if the goal of corporate management is to avoid non-compliance, it is challenging for most professional managers to establish an effective compliance system in such an environment.[59]

In an environment where most companies either have no compliance requirements or engage in formal compliance, managers are expected to prioritize

---

[53] Many enterprises expand in times of boom without proper awareness of compliance. This lack of compliance consciousness may lead into regulatory breaches and governance issues, thereby exposing the companies to legal risks. Miriam H. Baer, Governing Corporate Compliance, 50 B.C. L. Rev. 949 (2009).

[54] John Braithwaite, Corporate Crime Research: Why Two Interviewers Are Needed, 19 Sociology 136, 136-138 (1985).

[55] Cary Coglianese & David Lazer, Management-Based Regulation: Prescribing Private Management to Achieve Public Goals, 37 Law & Society Rev. 691, 691-730 (2003).

[56] Ulrike Malmendier & Geoffrey Tate, Does Overconfidence Affect Corporate Investment? CEO Overconfidence Measures Revisited, 11 Eur. Fin. Mgmt. 649, 655 (2005).

[57] Id.

[58] Todd Haugh, Nudging Corporate Compliance, 54 Am. Bus. L.J. 683, 685 (2017).

[59] Some managers are myopic, considering only performance during their working life. This type of thinking can result in strategies aimed to achieve gains in the short term at the expense of long-term sustainability and that might ultimately compromise the future growth and stability of the company. Christine Parker & Sharon Gilad, Internal Corporate Compliance Management Systems: Structure, Culture and Agency, in Explaining Compliance: Business Responses to Regulation 170-197 (2011).

company interests. In a worse circumstance, managers are forced to transfer interests from stakeholders to shareholders.[60] In such cases, non-compliance becomes a more cost-effective decision. Nevertheless, the improvement of compliance systems is a dynamic process that has been evolving for over forty years. With the close relationship between internal control mechanisms and economic development, the U.S. Department of Justice has expanded and strengthened the implementation of criminal law content to include corporate compliance within its jurisdiction. Additionally, research indicates that well-operating companies are more cooperative with mandatory disclosure obligations and other compliance matters, making the construction of an effective compliance system essential for companies.[61]

2. Key Milestones in the Evolution of Compliance Systems

Professors Christine Parker and Sharon Gilad believe that in the daily operation of corporate compliance systems, there are three key points that frequently participate in the construction of substantive corporate activities.[62] These points are not the only paths for interpreting corporate compliance systems, but they are one of the models that can effectively categorize and summarize current corporate compliance experiences. The three key points are: connecting structures, individuals, and compliance culture. The following sections will detail their roles in the operation of compliance activities:

For the development of professional compliance management, the connecting structures are composed of densely distributed knowledge and professional skills.[63] In an effective compliance system, these nodes consist of numerous specialized employees such as compliance professionals, environmental managers, and in-house lawyers. Companies delegate authority to these professionals, who use their expertise to categorize and solve issues, ultimately ensuring that each problem is addressed most professionally, leading to effective compliance management. In practice, for a company, these individuals who integrate knowledge and professional skills serve as the connecting nodes between the company and business opportunities. They use their abilities to ensure the effective operation of compliance activities. However, it is inevitable that despite their efforts to balance obtaining business opportunities with legal control, thesse professionals will often have to make decisions that either achieve the company's commercial objectives or ensure the effective implementation of legal norms. For instance, they may need to present compliance conflicts inherent in decisions directly to top management for resolution or involve external regulatory bodies in the compliance system's management to prevent top management from pushing through certain actions while ignoring compliance. Therefore, to clearly understand a compliance system model, it is essential to clarify the management attitudes held by its internal professionals, whether they are more politically inclined or conciliatory. This will directly influence the relationship between the compliance system and external controls, as well as the internal operations of the compliance system. Ultimately, it will have a significant impact on the company's long-term

---

[60] Mark R. DesJardine, Muhan Zhang & Wei Shi, How Shareholders Impact Stakeholder Interests: A Review and Map for Future Research, 49 J. Mgmt. 400, 407 (2023).
[61] Stephen Owusu-Ansah, Factors Influencing Corporate Compliance with Financial Reporting Requirements in New Zealand, 15 Int'l J. Com. & Mgmt. 141, 141-157 (2005).
[62] See Parker & Gilad supra note 9, at 170-197.
[63] Id. at 167-168.

commercial decision-making habits and objectives.

In U.S. compliance, whether agency of individuals within the company incorporates their personal philosophies into the establishment of compliance standards will significantly affect the operation of the compliance management system within the company.[64] Particularly, the agency of management commitment demonstrated by top individuals in daily affairs greatly influences the successful implementation of compliance management.[65] For example, a classic compliance incentive measure is determining the level of compliance behavior norms that company personnel must achieve to avoid internal accountability. Employees, in most cases, will choose to maintain the company's compliance system to avoid potential substantial financial penalties, thereby promoting effective compliance within the company. Additionally, during the recruitment of senior personnel, instilling the understanding of the high market operating costs associated with non-compliance before they join the company can lead them to effectively fulfill their fiduciary duties to the company and maximize corporate interests through effective compliance. Although building a compliance culture is not instantaneous, sustained policy pressure and continuous management can ultimately influence executives' and employees' understanding of the relationship between compliance concepts and business objectives, leading to the establishment of an effective compliance culture.[66] Furthermore, the influence of compliance culture extends beyond top management awareness. The feedback structure from the numerous employees within the company can also inversely affect the development and final implementation of the compliance system,[67] even adjusting top management's awareness. Moreover, the occurrence of sudden financial events can also highlight the need to improve the compliance system. In reality, companies without compliance often recognize the risks of non-compliance only after experiencing a crisis. This realization prompts top management and other personnel to advocate for the construction of a compliance system to mitigate future risks in different parts, including the decisions, internalization and implementation.[68]

In a compliance system, the timeliness and accuracy of information transmission significantly impact the system's operational efficiency. Additionally, how internal employees understand the compliance system contributes to the company's compliance culture.[69] Culture mediates between formal compliance systems and individual actions by influencing perceptions and shaping responses.[70] Particularly in terms of the distribution of compliance obligations, distributing them across every entity within the process is a more rational approach. Compliance is not solely based

---

[64] See Parker & Gilad supra note 9, at 178.

[65] The ideas and views of the top management will dominate the company's operating environment. From a strategic point of view, their vision and styles of decision-making not only set up the corporate culture and organizational priorities but also the overall direction of the business. See Id. at 179.

[66] Sharon Gilad, Accountability or Expectations Management? The Role of the Ombudsman in Financial Regulation, 30 Law & Policy 227, 227-253 (2008).

[67] See Parker & Gilad supra note 9, at 179.

[68] Id.

[69] Id.

[70] The environment in which the company operates will profoundly impact the governance of the company. The external and internal factors include the regulatory framework, market conditions, organizational culture, and stakeholder expectations that mold the governance practices and policies. Id.

on top-level management or the combination of professional personnel; it is actually the responsibility of everyone in the company. This means that constructing a compliance system must consider the functioning of lower-level employees and management models. Designing the compliance system according to the value orientation of these employees will make the process smoother. For example, establishing promotion-based incentives in the compliance system is one approach.[71] Offering cash rewards or promotions to lower-level employees who provide important compliance-related information that prevents significant company losses is another. In this model, information transmission within the company system becomes smoother and more effective, and employees will concentrate on collecting and analyzing effective information, which can be crucial when the company needs to comply. It is also essential to reasonably divide the responsibilities of employees. It should not be determined by the employee's level but rather guided by their obligations. This prevents the creation of ambiguity regarding duties and rights within the company culture, which could otherwise lead to distrust in the company's operations and hinder the normal functioning of the compliance system.

3. Compliance Obligations of Directors

According to Delaware law, directors will not choose to engage in unlawful actions that are profitable for the company.[72] New York law stipulates that directors are liable for illegal actions, even if those actions benefit the company.[73] In corporate governance, a significant proportion of companies attempt to comply with the law or intend to comply with the law. However, when faced with complex and potentially ambiguous legal interpretations, they may choose to circumvent the law to maximize their profitability. [74]Therefore, it is crucial to clearly define the compliance obligations of directors and the standards by which these obligations are judged.

Regarding compliance obligations, the primary duty model actually stems from fiduciary duty. The fiduciary duty, etymologically, is based on the trust a principal places in an agent to handle their affairs.[75] It is commonly applied in many areas involving agency relationships, from basic agency to partnerships and corporate governance.[76] Generally, fiduciary duty requires the agent to act in the best interest of the principal, especially when their own interests conflict with those of the principal.[77] In terms of compliance matters, the relationship between directors and the company is essentially a fiduciary one. This means that directors, based on their fiduciary duty,

---

[71] Ivo Hristov, Riccardo Camilli & Alessandro Mechelli, Cognitive Biases in Implementing a Performance Management System: Behavioral Strategy for Supporting Managers' Decision-Making Processes, 45 Mgmt. Res. Rev. 1110, 1111 (2022).

[72] See Marc S. Gerber, Edward B. Micheletti Peter & A. Atkins, Directors' Fiduciary Duties: Back to Delaware Law Basics, Skadden (February 19, 2020), https://www.skadden.com/insights/publications/2020/02/directors-fiduciary-duties.

[73] NY Bus Corp L § 719.

[74] The balance between compliance costs and non-compliance benefits has been struck. Firms do take the cost of compliance into consideration against their probable short-term gains for bypassing these rules. David B. Spence, The Shadow of the Rational Polluter: Rethinking the Role of Rational Actor Models in Environmental Law, 89 Calif. L. Rev. 917, 974-75 (2001).

[75] See Marc S. Gerber, Edward B. Micheletti Peter & A. Atkins supra note 33.

[76] Id.

[77] Id.

must endeavor to avoid commercial risks that could potentially harm the company, thereby protecting the company's interests.

In the United States, there is a classic explanation of fiduciary duty, which states that directors are responsible to the shareholders or the company.[78] These fiduciary duties include four main components: the duty of loyalty and the duty of care. Professor Black believes that two additional duties should be included: the duty of disclosure and the duty of extra care when the company becomes a takeover target. [79]This article will also incorporate these two duties into the comparative legal discussion and will discuss them one by one:

The fiduciary duty of loyalty is paramount, requiring decision-makers to prioritize the company's interests over their own, avoiding self-dealing or dual agency.[80] For many companies, especially smaller ones, completely prohibiting self-dealing is impractical due to limited transaction channels.[81] If a director can bring profitable opportunities to the company, such actions may not need to be prohibited.[82] In the U.S., a conflict of interest transaction will be recognized by the court if it was conducted under fair negotiation, with disinterested directors' approval, unless shareholders can prove the transaction was unfair to the company.[83] Independent directors typically assess such transactions, and even if the company might bear the damages, they often restrict interested directors' voting to protect their and the company's credibility from public criticism, which is the outcome of a cost-benefit calculation.[84]

The duty of care, as the second type of fiduciary duty, typically applies in situations where there is no conflict of interest.[85] This duty requires directors to maintain a prudent attitude and make well-considered decisions.[86] The duty of care does not require directors to make perfect decisions; rather, it demands that directors be present, act prudently, and make business decisions that are not completely irrational.[87] In essence, directors must remain diligent and relatively reasonable in their actions towards the company.[88]

---

[78] Id.
[79] Bernard Black, The Principal Fiduciary Duties of Boards of Directors, OECD (April 4, 2001), https://www.oecd.org/daf/ca/corporategovernanceprinciples/1872746.pdf.
[80] Id.
[81] Id.
[82] In actual operations, it is not uncommon for company directors to leverage their personal connections to benefit the company. This practice can provide strategic advantages, such as gaining access to new markets, securing favorable contracts, or fostering beneficial partnerships. While this can be advantageous for the company, it also requires careful governance to ensure that such practices align with ethical standards and do not lead to conflicts of interest or undermine corporate integrity. Id.
[83] Id.
[84] Wei Jiang, Hui Wan & Shan Zhao, Reputation Concerns of Independent Directors: Evidence from Individual Director Voting, 29 Rev. Fin. Stud. 655, 670 (2016).
[85] Bernard Black supra note 40.
[86] Id.
[87] Id.
[88] Julian Velasco, The Diminishing Duty of Loyalty, 75 Wash. & Lee L. Rev. 1035 (2018).

The duty of disclosure will be fiduciary in nature is doubtable,[89] but it is already mandated by securities law for publicly traded companies, and it is also a crucial compliance obligation.[90] Directors must firmly fulfill their disclosure obligations to shareholders in the following two situations: when shareholders are required to vote and when the company has just completed a conflict-of-interest transaction. Regarding shareholder voting, shareholders need to analyze and vote based on the actual situations they are aware of; thus, they have the right to know all information that might affect the voting results. As managers of the company's daily affairs, directors naturally become the subject of this disclosure. Regarding conflict-of-interest transactions, two main dimensions need consideration. First, after disclosing all details of the transaction to shareholders, as previously mentioned, shareholders can block the completion of the transaction if they can prove it is not entirely fair to the company. Second, the information might reveal that directors have violated their duty of loyalty, providing grounds for shareholders to sue the directors.

Directors also have a special additional duty of care.[91] When a company becomes a takeover target, the board of directors may face a situation where their interests are significantly affected.[92] After the company is acquired, all members of the board may lose their management positions, even if they assume certain management roles in the new company, these roles will differ from their previous powers as directors.[93] Therefore, in this situation, all directors become interested parties, and whether to approve the acquisition becomes a conflict-of-interest transaction.[94]

At this time, directors must fully disclose this information to the shareholders of the company, and the decision on whether to accept the acquisition should ideally be made by the shareholders themselves. This prevents directors from making decisions that violate their duty of loyalty due to their conflicting interests.

4. Application of the Business Judgment Rule

In the United States, the business judgment rule effectively protects directors and senior executives from excessive scrutiny regarding their compliance obligations.[95] This protection is based on several key points: First, courts should not second-guess business decisions after the fact because unfavorable outcomes in business investments can result from various factors, and poor management decisions might be just one of

---

[89] Bronwen Nosworthy, A Directors' Fiduciary Duty of Disclosure: The Case(s) Against, 39 Univ. N.S.W. L.J. 1389, 1408 (2016).

[90] Matthew C. Turk and Karen E. Woody, The Leidos Mixup and the Misunderstood Duty to Disclose in Securities Law, Harvard Law School Forum on Corporate Governance (July 21, 2017), https://corpgov.law.harvard.edu/2017/07/21/the-leidos-mixup-and-the-misunderstood-duty-to-disclose-in-securities-law/.

[91] When a company becomes a target for acquisition, directors, due to their unique positions, will bear additional responsibilities. Bernard Black supra note 40.

[92] Id.

[93] Id.

[94] Id.

[95] Ivan R. Gutierrez, The Business Judgment Rule: A Shield and Sword, ABA (July 22, 2019), https://www.americanbar.org/groups/litigation/resources/newsletters/business-torts-unfair-competition/business-judgment-rule-shield-and-sword/.

them.

Second, unfavorable outcomes are among the business risks that shareholders should have anticipated from the outset. Third, many business decisions involve risks that can yield significant benefits for the company, although some might result in losses. If directors were held liable for every adverse outcome of risky business decisions, they would be less inclined to take such risks, which could reduce the number of potentially beneficial decisions made. Removing directors' discretionary power could result in overly cautious decisions that appear to meet compliance goals but ultimately hinder the company's long-term development, which compliance itself aims to support. So crucially, the court's inquiry does not focus on the substance, not even to a minimal rationality standard, but rather on whether the decision ultimately made was a 'rational result of' the decision-making process that was actually carried out.[96]

In essence, the business judgment rule provides directors with a certain level of immunity when making risky decisions, ensuring they can act in the company's best interest without the constant threat of personal liability for adverse outcomes.

B. The Compliance System Framework Under Japanese Law

In Japan, compliance systems are integrated into broader internal control systems. Although the Japanese Companies Act does not mandate internal control systems, directors are expected to adhere to relevant laws and company regulations. The duty of care and loyalty form the legal basis for these systems. Key elements of Japan's internal control system include a control environment, risk assessment, control activities, information and communication, and monitoring activities. These components ensure that corporate operations are conducted ethically and legally, with directors playing a crucial role in maintaining compliance and addressing risks through a structured and proactive approach.

1. Legal Basis

In Japan, compliance systems are actually included in internal control systems, also known as risk management systems. In fact, the Japanese Companies Act does not mandate the establishment of internal control systems. It only requires directors to adhere to relevant laws, [97]regulations, and the company's articles of incorporation when performing their duties, [98] without a system mandating the board to fulfill corresponding internal control responsibilities. [99] However, if a company has established an Audit Committee, then it must construct an internal control system.

According to previous Japanese case law, the board of directors, as the executors

---

[96] Lyman Johnson, The Modest Business Judgment Rule, 55 Bus. Law. 625, 650 (2000).
[97] In Japan, setting up internal control systems is a judgment call of companies. This kind of independence enables every company to design the proper internal controls over their operational needs and business environments. Kaisha-hō [Companies Act], Law No. 86 of 2005, art. 416, para. 1 (Japan).
[98] Id.
[99] Id.

of company affairs, has the obligation to establish an internal control system. The legal foundation for constructing this internal control system lies in the duty of care of a good manager. [100]The legal basis for the internal control system can include the duty of prudence, the duty of care, the duty of supervision, or the duty of monitoring.

2. Basic Elements of the Internal Control System

The basic elements of internal control are the necessary components for achieving internal control objectives. Referring to the "Standards for Evaluation and Audit of Internal Control over Financial Reporting" issued by the Financial Services Agency of Japan (hereinafter referred to as the "Internal Control Standards"), the basic elements of internal control mainly include six key components:[101]

The first key component is control environment that refers to the values held by an organization and its basic personnel and work structure systems.[102] It is the key component of the internal control system, determining the attitudes towards control among various parts of the organization and within the internal control system of each individual. It forms the foundation for the other elements.[103] The organizational culture often reflects the intentions and attitudes of the Chief Executive Officer towards the internal control system and other company systems. Additionally, the value standards held by the organization and the basic systems of the internal control system will similarly influence the attitudes of personnel within the organization towards the internal control system. Generally, the usual aspects of the control environment include but are not limited to: a. Organizational integrity and ethics. b. Management's intentions and attitudes. c. Management policies and strategies. d. Division of functions between the board of directors and auditors. e. Organizational structure. f. Allocation of rights and responsibilities. g. Human resources policies and procedures.[104]

The second key component is risk assessment that refers to the process of identifying, analyzing, and evaluating events that could potentially or have already impacted the achievement of organizational objectives.[105] Specifically, risks can be analyzed from two levels: external factors and internal factors. External factors include natural disasters, external illegal acts, market competition changes, foreign exchange impacts, etc. Internal factors include information system failures, top management decision failures, etc.[106] Due to the specific business characteristics and other factors, the approach to risk assessment will vary. The general process is as follows: First is risk identification. The first step in the risk assessment and response process is to establish

---

[100] 東京高等裁判所 [Tokyo High Ct.] May 21, 2008, 1281 判例タイムズ [Hanrei T.] 274 (Yakult Case) (Japan.).

[101] Financial Services Agency (金融庁), Standards for Evaluation and Auditing of Internal Control over Financial Reporting (財務報告に係る内部統制の評価及び監査の基準), Financial Services Agency (金融庁) (December 6, 2019), https://www.fsa.go.jp/singi/singi_kigyou/kijun/20191206_naibutousei_kansa.pdf (Japan).

[102] Id.

[103] Id.

[104] Id.

[105] Within an internal control system, identifying risks is of paramount importance. Effective risk identification forms the foundation for developing appropriate mitigation strategies and safeguards. Id.

[106] Id.

how to correctly identify risks.[107] Since risks exist in various business processes within the company, risk identification should be a continuous and phased process to more effectively discover events that may impact the achievement of organizational objectives. The second step is risk classification. Risks can be categorized as historical risks and future risks, as well as company-wide risks and business process risks. The third step is risk analysis. Each risk has a different impact on the company. Generally, companies take corresponding actions against significant risks. This step involves identifying the potential impact of each risk on the company. The fourth step is risk response.[108] I hold the standpoint that here are four main response measures: a. Risk avoidance: This is chosen for situations where the risk impact is very high or difficult to control. b. Risk mitigation: This involves taking corresponding actions to establish internal controls that reduce the impact of the risk. c. Risk transfer: This involves transferring all or part of the risk outside the company to reduce its adverse impact, such as purchasing relevant insurance or conducting direct risk hedging transactions. d. Risk acceptance: This means taking no action against the risk. This measure is a result of balancing the benefits and costs behind it. If the cost of addressing the risk exceeds the benefits, the risk should be accepted to gain more benefits, and vice versa.

The third key component is Control activities that refer to a broad range of policies and procedures implemented to ensure that management's directives are properly executed.[109] For example, the separation of authority and responsibilities: management will clearly define the boundaries of power and responsibility for each process owner, and establish a system based on this to ensure that the responsible person performs their duties within appropriate limits, thereby minimizing the risk of fraud and errors.[110] Additionally, these responsibilities should be distinctly separated. This separation not only facilitates accountability but also ensures mutual checks and balances among responsible parties, thereby improving the overall operational efficiency of the system. Furthermore, as mentioned in the risk assessment above, control activities are closely related to it. A company must ensure the completeness of control activities to implement correct measures to address risks. In practice, the emergence of risks can indeed stem from the inadequacies within the control activities themselves. Regarding the monitoring activities to be discussed below, the completeness of control activities can also establish a more efficient monitoring system. Therefore, ensuring appropriate control activities within the internal control system is crucial.

The fourth key component is information and communication systems that refer to the processes that ensure necessary information is identified, analyzed, processed, and effectively communicated within and outside the company.[111] These systems

---

[107] This process involves systematically analyzing all potential threats to the organization's operations, financial integrity, and compliance obligations. Id.

[108] Id.

[109] Id.

[110] In a company, defining the scope of duties and powers of management personnel is extremely important. Clearly delineating these responsibilities ensures that managers understand their roles and the extent of their authority, which is crucial for effective governance and decision-making. Id.

[111] Information can be said to be the engine that drives corporate governance. Perfect information flow within a company ensures that there is an availability of relevant, correct, and timely information to all levels of the organization, enhancing transparency, accountability, and the ability to make informed decisions. Id.

enable timely and appropriate capture, identification, and transmission of information required by all relevant internal and external parties to perform their duties. Whether manual or automated, these systems handle the processing and transmission of information.[112] More importantly, the information needs to be not only transmitted but also understood correctly by the recipient and shared with the necessary individuals within the company. The primary process of this system allows the company to identify genuine and valid information. If the information is deemed necessary for the company, it is incorporated into the information system.[113] Additionally, this system typically includes a whistleblowing mechanism to enhance the collection of external information. Although this is not a conventional communication channel, it serves as a mechanism for information transmission and supervision, allowing all members of the company to convey information to the board of directors, auditors, or in some cases, to external contacts like lawyers.[114] When introducing a whistleblowing system, the board should take measures to ensure its effective operation. A key aspect of this is establishing an effective whistleblower protection mechanism to prevent retaliation against whistleblowers, thereby maintaining the overall operation of control activities. This information management system is closely linked to other elements. For example, when a new policy is established in the control environment, this information is communicated to the appropriate personnel within the organization and understood accurately to facilitate subsequent risk assessment and control activities. If significant deficiencies are identified in control activities or monitoring processes during operation, this information will also be promptly and accurately communicated to the board of directors to innovate the control environment and implement appropriate control activities. Therefore, it is crucial to ensure that the company's organizational structure related to information is rational and to guarantee a high-quality information communication system.

The fifth key component is Monitoring, which is the continuous process of assessing whether the internal control system is functioning effectively.[115] This includes routine monitoring integrated into business operations and independent evaluations separate from business activities. Both types can be conducted independently or in combination.[116] Routine monitoring involves the continuous review and assessment of the internal control system's effectiveness through procedures integrated into regular business operations.[117] For instance, in financial reporting, this might involve periodic or ad-hoc supervision of accounts receivable balances by management. Any discrepancies identified are promptly analyzed and corrected, making this process effective in verifying related financial information. Independent evaluation, on the other hand, can be conducted by three main entities: a. The Board of Directors: Given that the board determines the basic control environment and oversees the execution of company affairs in daily operations, it is effective for the board to conduct independent evaluations. b. The Supervisory Board: The supervisory board,

---

[112] Id.
[113] Id.
[114] Id.
[115] Monitoring of commercial decisions and member behavior within the company is also very important within an internal control system. It ensures that the decisions made are in line with the corporate policies and strategic objectives of the organization and directs proper ethical conduct that answers to regulatory standards. Id.
[116] Id.
[117] Id.

established within the company, is responsible for overseeing daily business operations. Its role in independent evaluation helps to effectively improve the internal control system. c. Auditors: Auditors can audit the performance of the board of directors and the supervisory board and conduct investigations using relevant personnel.[118] This ensures the timely detection of internal control deficiencies. d. Following these monitoring activities, a report on the internal control system's monitoring results should be promptly generated and communicated through a comprehensive information transmission mechanism within the internal control system.[119] Effective adjustments should be made in a timely manner based on this report.

The sixth key component is information technology readiness that refers to the level of information technology required to achieve the appropriate functioning of the first five factors in business operations, based on the company's pre-established policies and procedures.[120] Most effective information management systems today utilize information technology, and there is a significant difference in information processing capabilities among various technologies. For example, the latest product from OpenAI, ChatGPT, possesses an extremely high level of information processing technology due to its advanced artificial intelligence module. Integrating such advanced information processing technology into a company's internal control system can have a substantial positive impact, significantly reducing the costs associated with the relevant system. Although AI processing cannot replace humans as the ultimate source of information generation, it can process and learn from the continuously generated information within the company. This makes the information system more aligned with the needs of the inquirers, enhancing the control environment of the information system.[121]

3. The Duties of Directors

In Japanese law, the duties of directors are closely related to their supervisory obligations. Regardless of whether a board of directors is established within the company, directors responsible for business execution naturally bear the duty of supervising their respective areas. With the increasing emphasis on corporate compliance, Japanese scholars believe that the supervision by directors should not be limited to direct supervision by superiors but should also encompass the supervision of company employees and overall business operations.[122] The board of directors is the core of the company's daily operations, directly influencing changes in the control environment of the internal control system and shaping the company's future direction.[123] Whether they are employees, middle management, or senior executives

---

[118] Id.
[119] Id.
[120] Modern models of corporate governance can leverage technology for governance to significantly enhance effectiveness. Advanced technological devices permit more accurate analyses of the data, smoother processes, and better decision-making. Id.
[121] Holden Thorp, ChatGPT Is Fun, but Not an Author, 379 Science 313, 313 (2023).
[122] Liang Shuang (梁爽), Mei, Ri Gongsi Fa Shang de Dongshi Hegui, Neikong Yiwu Ji Qi Dui Woguo de Qishi (美、日公司法上的董事合规、内控义务及其对我国的启示) [Directors' Compliance and Internal Control Obligations in U.S. and Japanese Corporate Law and Their Implications for China], 2 Zhongwai Faxue (中外法学) [Peking University L.J.], 530 (2022).
[123] The board of directors may go on to have a great influence on the long-term business decisions of a company. Their strategic guidance and oversight set the vision, mission, and total direction for a company. Kaisha-hō [Companies Act], Law No. 86 of 2005, art. 348, para. 1 (Japan).

such as directors, all are bound by the control environment. Therefore, directors inherently have corresponding supervisory obligations and other organizational duties over all entities affected by the control environment.

In Japanese law, the legal duties of directors are divided into the duty of care and the duty of loyalty. The duty of care requires directors to act with caution while supervising and organizing company management activities. For instance, the duty to report, as stipulated in Article 357 of the Japanese Companies Act, mandates that if a director discovers a fact that may significantly impact the company, [124] they must promptly report this fact to the shareholders and the board of auditors. The duty of loyalty obliges directors to act in the best interests of the company and its shareholders when supervising and organizing company management activities. [125] This duty requires directors to prioritize the company's and shareholders' interests over their personal interests.[126]

4. Application of the Business Judgment Rule

In Japan, the business judgment rule serves as a decision-making framework. It generally allows directors and executive officers the discretion to make business-related decisions, provided that the content and process of these decisions are not unreasonable. Under this rule, such decisions would not be considered as violations of the duty of care. Typically, the business judgment rule is also applicable to the obligations related to the construction of internal control systems.[127] Directors within the company structure are regarded as management experts. Shareholders entrust the directors with management authority, allowing them to utilize their management skills and fulfill their responsibilities to ultimately maximize the company's interests. Throughout this process, directors are afforded the utmost respect, and retrospective evaluations of their decisions for potential violations of the duty of care are not permitted.

Directors should take bold risks to maximize company profits, but excessive liability for damages in case of failure can lead to bankruptcy, causing directors to avoid necessary risks, thus impacting company benefits. [128] Legal policy should remove factors causing directors to hesitate by pre-limiting compensation amounts to a lower amount or zero, ensuring directors can take necessary risks without fear of personal financial ruin.[129] While profits from successful risk-taking benefit the company, losses should not be solely imposed on directors; the company should share both benefits and

---

[124] Hisahisa Uedaya (恒久 上田谷), Limitation of Directors' Liability to the Company [取締役の対会社責任の制限], 39 Tsukuba L.J. [筑波法政] 145, 149 (2005) (Japan).
[125] Id.
[126] Id.
[127] It simply means that even within the operation model of an internal control system, the division of duties among the internal structures can also be based on the business judgment rule. In this principle, the internal control system simply makes sure that the decision-making process is made in good faith, with due care, and in a manner believed to be in the best interest of the corporation. TOSHIO SAKAMAKI (酒巻俊雄) & SETSU TATSUTA (龍田節), ARTICLE-BY-ARTICLE COMMENTARY ON THE COMPANIES ACT, VOLUME 5 (逐条解説 会社法 第 5 巻) 362 (2011).
[128] Hisahisa Uedaya Supra note 63, at 150.
[129] Id.

risks.[130] Legal accountability for directors' negligence should only be pursued when there is clear evidence. Fairly limiting liability encourages directors to perform effectively, preventing the loss of valuable director talent. [131] A balance is needed between safeguarding company interests and motivating directors to take necessary risks.[132]

C. Horizontal Comparison

This chapter focuses on the organizational frameworks and the duties of directors and in the compliance systems of the United States and Japan. Common elements in the overall framework of compliance organizations can be summarized as follows: having an upper organizational intent, compliance organization members, information exchange mechanisms, supervision mechanisms, remedial mechanisms, and the policies corresponding to these mechanisms: First, having an upper organizational intent determines the compliance environment of the company. The form of compliance and its direction must be established through the expression of the upper organizational intent. Establishing a good upper organizational intent is crucial to the effectiveness of company compliance, akin to the "constitution" within a company's compliance system. All major compliance activities of the company will be based on this intent. Second, compliance organization members are the specific implementers of the compliance system. This includes the board of directors, the supervisory board, independent auditors, external directors, and internal employees. The specific roles they undertake in the compliance process need to be clearly defined by the company. Third, the information exchange mechanism is one of the most critical components of the system. In the compliance system, the transmission of information is responsible for conveying risk information promptly to the management and executive levels. Effective and efficient information transmission is the foundation of an effective compliance system. If information transmission is hindered, the compliance system will ultimately fail to resist risks, regardless of how well other parts of the system operate. Fourth, the supervision mechanism is closely related to the information exchange mechanism. While the information exchange mechanism collects and classifies risk information, the supervision mechanism analyzes and reviews all information and personnel, promptly reporting factors that might affect the effective operation of the compliance system to the relevant compliance officers through the information exchange system. This mechanism typically needs to be established as an independent system to prevent undue influence from interested parties. Fifth, the remedial mechanism is a dynamic system that addresses risk issues identified by the supervision system, preventing the escalation of harm. It involves preemptive remedies before risk occurrence, corrective measures during risk occurrence, and solutions after risk occurrence. The effectiveness of the compliance mechanism is judged by whether risks have impacted the company or have

---

[130] Id.

[131] It would provide greater opportunities for the company's directors to perform effectively if they were given greater autonomy. Greater autonomy will permit directors to exploit their expertise, make rapid strategic decisions, and innovate without obtrusive controls. Id.

[132] Excessive autonomy given to directors can go against the interests of the company; therefore, it is important that regulatory design provides an effective balance. That means the liberty for making strategic decisions has to be available to the directors, but it needs to go with checks and balances that obviate its abuse. Id.

been promptly mitigated by the company.

Regarding the duties of directors, senior management within the company should owe fiduciary duties to the company, including the core duties of loyalty and diligence. These duties require directors to prioritize the company's interests and handle company affairs with a cautious attitude.

Concerning compliance responsibilities, both countries have introduced the business judgment rule as a core standard for determining whether directors and other management personnel have breached their duties. Under this standard, directors have the discretion to make business decisions, ensuring an effective balance between benefits and risks, rather than avoiding all risks in their decision-making.

The understanding of the business judgment rule in the United States and Japan is not entirely the same. In the United States, the principle of the business judgment rule is that the content of business decisions is not subject to post hoc review by the courts. The perspective is that if the presumed actions of directors are made in the best interest of the company, the courts will not scrutinize the management decisions after the fact.[133]

In contrast, Japanese courts, when applying the business judgment rule, analyze and review both the process of forming the business judgment and the content of the judgment. The Supreme Court of Japan, in its 2010 case No. 2091, ruled that in the process of considering acquisitions, directors should take into account the valuation of shares and the appropriateness of the acquisition when making decisions. As long as the process and content of their decision are not grossly unreasonable, it will be interpreted as not violating the duty of care expected of directors. This indicates that the Japanese Supreme Court's interpretation of the business judgment rule involves examining whether directors' business decisions meet the standard of care expected of a prudent manager. In the United States, the application of the business judgment rule is understood as requiring verification of relevant facts before making a business judgment. As long as directors make decisions that are suitable for the company based on these facts,[134] they are not considered to have violated their duty of care, even it cannot be suitable in the perspective of Delaware's court.[135]

IV. ESTABLISHING AN IDEAL COMPLIANCE SYSTEM MODEL

This section proposes a dual-level compliance concept tailored to China's unique legal environment, integrating both private and public interests. At the private level, companies focus on profitability and efficient operations, balancing shareholder value and compliance costs. At the public level, compliance aims to meet societal

---

[133] Hindsight judgment is impermissible under the business judgment rule in US. The rule protects the directors from unfair criticism by demonstrating that decisions, upon hindsight, appeared at that time to be reasonable and informed, even though the outcome may not turn out favorable. MITSUO KONDO (近藤光男), KEIEI HANDAN GENRI (判例法理 経営判断原則) 12 (2012).

[134] Bernard S. Sharfman, The Importance of the Business Judgment Rule, 14 NYU J.L. & Bus. 27, 49 (2017).

[135] Id.

expectations, ensuring market stability and enhancing social welfare. The proposed model emphasizes robust information transmission, whistleblower protections, and temporary management systems to enhance effectiveness. By clearly defining compliance duties for directors, supervisors, and employees, the model aims to create a comprehensive and adaptable compliance framework that supports sustainable and ethical corporate governance.

A. Duality of Compliance Concepts

An effective local compliance concept for China should encompass two levels: the private level and the public level. At the private level, companies need to consider their primary goal of profitability, which requires balancing the value of private interests. This involves ensuring that the company operates efficiently and profitably while adhering to compliance standards that protect its interests. At the public level, the focus is on the societal expectations of companies.[136] A well-functioning company should not only ensure the realization of its own interests but also contribute to the overall social value.[137] From the perspective of compliance concepts, this might sometimes involve sacrificing certain company interests to ensure the stability of the overall market transactions and enhance societal welfare. This dual approach ensures that companies not only thrive economically but also uphold their responsibilities towards broader societal goals.

1. Consideration of Private Interests

The consideration of private interests encompasses two dimensions. First, it involves the long-term development needs of the company's value. Second, it addresses the needs of the company's shareholders. Particularly under the theory of shareholder primacy, a company should adopt effective compliance plans to prevent behaviors that could harm its value, thereby ensuring capital stability and protecting the stability of shareholder interests, which have been converted into equity.

Furthermore, from a long-term perspective, the establishment of a compliance system can positively enhance the social reputation of the company and its shareholders. This improved reputation can, in turn, positively influence the company's competitiveness in various business opportunities, reduce transaction costs, and increase transaction efficiency. Therefore, when designing specific compliance concepts, it is essential to ensure the rationalization of transaction costs, the rationalization of compliance costs, and the maintenance of transaction opportunities as part of the compliance concept for private interests:

a. Rationalization of Transaction Costs

In the course of daily transactions, companies should avoid engaging in transactions where the estimated transaction costs significantly exceed the estimated

---

[136] David S. Ruder, Public Obligations of Private Corporations, 114 U. Pa. L. Rev. 209, 211 (1965).
[137] Id.

transaction benefits.[138] In practice, companies face limitations in judging the opportunities of transactions and cannot always directly determine the exact benefits and costs. Transactions are often evaluated before they are concluded. Transactions with relatively high costs are not necessarily excluded by the compliance concept but are assessed based on the potential estimated benefits. If the estimated benefits exceed the estimated costs, the transaction can proceed. However, this does not guarantee absolute profitability. If directors approve such a decision and the company incurs a loss from the transaction, the company cannot demand compensation from the directors solely based on this loss.[139]

b. Rationalization of Compliance Costs

Ensuring the rationalization of compliance costs is crucial for maintaining the financial health and operational efficiency of a company. Effective compliance programs should be designed to minimize unnecessary expenses while maximizing the benefits of compliance. This involves implementing streamlined processes, leveraging technology, and training employees efficiently to reduce the costs associated with compliance. By doing so, companies can avoid excessive spending on compliance measures that do not proportionately contribute to risk mitigation or operational efficiency. Rationalized compliance costs ensure that the company maintains its competitive edge without sacrificing legal and regulatory adherence. Furthermore, a cost-effective compliance strategy enhances the company's ability to invest in growth opportunities and innovation, ultimately contributing to its long-term success and stability.

c. Maintenance of Transaction Opportunities

Maintaining transaction opportunities is vital for a company's growth and competitive positioning, which can be a risk and needed be safeguarded.[140] So a robust compliance framework should not hinder the ability to engage in beneficial transactions but rather support it by creating a reliable and trustworthy business environment. This involves ensuring that compliance measures are flexible and adaptable to changing market conditions and opportunities. By fostering a culture of compliance that emphasizes the importance of ethical behavior and transparency, companies can build strong relationships with partners and stakeholders, thus preserving and enhancing transaction opportunities. Effective compliance systems help in identifying and managing risks proactively, allowing the company to seize profitable opportunities while mitigating potential downsides.[141] This balanced approach ensures that

---

[138] In the course of business decision-making, among the expected factors to be dealt with seriously are risks, costs, and benefits. The decision-maker has to analyze carefully the possible hazards, financial implications, and expected returns. Oliver E. Williamson, Transaction Cost Economics, 1 Handbook of Industrial Organization 41, 42-44 (1989).

[139] L. McMillan, The Business Judgment Rule as an Immunity Doctrine, 4 Wm. & Mary Bus. L. Rev. 521, 569 (2013).

[140] Jeffrey H. Dyer, Effective Interim Collaboration: How Firms Minimize Transaction Costs and Maximize Transaction Value, 18 Strategic Mgmt. J. 535, 537 (1997).

[141] This points to developing an effective compliance system that mitigates risks but maximizes benefits. The law and other regulatory standards help business organizations avoid penalties and reputational damage if complied with, thus safeguarding operations. Id.

compliance supports rather than stifles the company's strategic initiatives and market responsiveness.

2. Public Interest

The purpose of corporate compliance extends beyond considering the interests of a specific company and its shareholders. As an organization, a company is also a part of the nation and should recognize its role as a "citizen." As a citizen, a company must ensure that its operations, especially compliance processes, do not harm other companies or stakeholders while safeguarding its own risk avoidance.

According to classical Western political thought, particularly the individualism-based liberal tradition, a company as a citizen has a passive citizen image. Although its core purpose is to protect its own rights, the criteria for judging whether a company is a good citizen include paying taxes, adhering to laws, and not violating regulations. On the other hand, the republican tradition views citizenship as the qualification to participate in public discourse. The ideal good citizen prioritizes public affairs over private matters and actively contributes to the construction of public affairs to demonstrate its value. Thus, under this tradition, corporate compliance inherently involves greater public interest.[142]

Therefore, in designing specific compliance concepts, ensuring overall market safety and stability, protecting creditor interests, and safeguarding employee interests should be embedded within the compliance system.

B. Typification of Corporate Compliance Rules

Typification of Corporate Compliance Rules is needed in China,[143] thus this section categorizes rules on directors' compliance duties into three main areas: decision-making, supervision, and remediation. Directors must gather and utilize relevant information to make informed decisions that benefit the company while minimizing societal harm. They are responsible for overseeing compliance-related activities within the company, ensuring adherence to regulations, and monitoring other directors and employees. When compliance issues arise, directors must take corrective actions, establish temporary management systems, and design effective remediation standards. Additionally, the section outlines the compliance obligations of supervisors

---

[142] It is intrinsically conceived that a compliance system shall serve not only the interest of shareholders in a company but also protect interests of society. While guiding itself with the tenets of the rule of law and ethical guidelines, it makes sure that business activities will have a positive social impact, be environmentally friendly, and respectful toward human rights. Wan Jianlin (万健琳) & Yang Songlei (杨松雷), Gongmin yu Guojia Hexie Guanxi de Goujian—Gonghezhuyi Lilun Yiyun Zai Wajue (公民与国家和谐关系的构建—共和主义理论意蕴再挖掘) [Constructing Harmonious Relationships between Citizens and the State—Further Excavation of Republican Theory], 4 Shandong Shehui Kexue (山东社会科学) [Shandong Soc. Sci.] 122, 126 (2015) (China).

[143] Dong Zhang (张栋) & Ke Li (李轲), Zhuti Lun Shiyu Xia She'an Qiye Hegui Youxiaoxing Shencha de San Zhong Moshi (主体论视域下涉案企业合规有效性审查的三种模式) [Three Models of Reviewing the Effectiveness of Corporate Compliance Involved in Cases from the Perspective of Subject Theory], 3 Jiaoda Faxue (交大法学), 129, 136 (2024) (China).

and other key personnel, such as compliance officers and information system members, emphasizing the importance of information accuracy, timely reporting, and whistleblower protections. This comprehensive approach ensures effective compliance management, balancing internal governance with external legal requirements.

1. Directors' Compliance Duties

Directors' compliance duties can be divided into the duty of loyalty and the duty of diligence within the realm of compliance.[144] In the context of compliance affairs, there are three main categories: decision-making, supervision, and remediation. This method allows for a typology of directors' compliance duties, as detailed below:

a. Decision-Making Duty

The compliance decision-making duty refers to the compliance obligations that directors must fulfill when making compliance decisions. These obligations generally involve two levels: the collection and utilization of decision-making information and the decision-making process itself.

Regarding the former, directors are obliged to be informed of information related to the decision-making content, and for important information, they should undertake a more rigorous duty of examination. For instance, when a company provides external guarantees, directors must ensure that the transaction has been approved by the shareholders' meeting. If such a decision has not been made, the directors can refuse to proceed with the transaction during the information collection stage.

As for the latter, the duty of decision-making does not strictly prohibit directors from making any decisions that might expose the company to risk. Rather, it requires them to make reasonable judgments based on the current operating conditions of both the company and the transaction partner, balancing the benefits and risks associated with the transaction. As long as directors effectively understand the information at this stage and make rational judgments in the best interest of the company based on that information. According to business judgement rule, which is simply a policy of judicial non-review.[145] In this case, they are deemed to have fulfilled their decision-making duties, court won't do the second guess.[146] Moreover, the decisions made by directors in the process of establishing or revising compliance systems are equally subject to the compliance decision-making obligations.

The determination of reasonable judgment requires consideration of two aspects: benefit to the company and minimal harm to society:

---

[144] In China, the new Company Law adopts a classic bifurcation strategy in designing the duties of directors. Jianwei Li (李建伟) & Kexin Ma (马可欣), Xinyi Yiwu Tixihua Sheji de Gongsi Fa Fang'an (信义义务体系化设计的公司法方案) [A Corporate Law Proposal for the Systematic Design of Fiduciary Duties], 5 Henan Shehui Kexue (河南社会科学), 36, 40 (2024) (China).

[145] Lyman Johnson supra note 96, at 632.

[146] Id.

Firstly, the benefit to the company means that the director's ultimate goal in making a decision is to help the company gain profits.[147] These profits need further definition and should be analyzed in conjunction with the company's long-term business plans. For example, if a company's long-term operational goal is to gather funds to expand production, and a current transaction can bring profits but requires a substantial one-time investment that affects the company's capital accumulation, the director should decide against the transaction to align with the company's long-term operational plan.

Secondly, minimal harm to society means that the director's decision should not seriously harm societal interests. In reality, it is impossible to ensure that every transaction completely aligns with societal interests, especially in situations where a company's actions might result from a balance of interests, ultimately sacrificing some societal interests for the company's benefit. According to the business judgment rule, this can also be a reason for a director's exemption from liability.

For decision-making, the ultimate aim is to apply these decisions in social practice. When there is a significant conflict between legislative regulations and social practices, directors may indeed have to make decisions that sacrifice the societal interests underlying those regulations in favor of aligning with the current social environment.

b. Compliance Supervision Obligation

The compliance supervision obligation requires directors to oversee compliance-related matters during the company's operations. This means that directors are not only responsible for ensuring their own decisions comply with regulatory requirements but also have a supervisory duty over the decisions of other directors and other compliance-related entities within the company.

To effectively fulfill the compliance supervision obligation, it should be constructed on two levels: establishing an effective information transmission and reporting system, and designing effective compliance supervision standards.

Establishing an effective information transmission and reporting system should be the core duty within this obligation. The effectiveness of compliance supervision directly depends on how information is obtained. In companies lacking a robust compliance system, there typically exists an unstable or non-existent information reporting system. A mature system should be independent and possess high-level access to company information. Independence ensures that information acquisition during supervision is not restricted by interested senior executives. High access permissions help the system obtain comprehensive information. If the system lacks such permissions, information can be bypassed through high-level communication, thereby

[147] This means that, in the case of a conflict of interest, directors are bound to give priority to the concerns of the company. They must step back from personal gains or pressures from elsewhere and make a decision that will benefit the company and the stakeholders. Lyman Johnson, Unsettledness in Delaware Corporate Law: Business Judgment Rule, Corporate Purpose, 38 Del. J. Corp. L. 405, 411 (2013).

restricting the full functionality of the reporting system.

Directors, supervisors, and employees' execution of compliance supervision duties require the company's standardized guidelines. As the core of company affairs management, directors have the obligation to design regulations on how supervision should be conducted. The standards should clarify the supervision obligations for each department, identify the responsible individuals within various departments (such as finance, design, logistics) and clarify the primary entities responsible for supervision within the board of directors and the board of supervisors.

Moreover, protection and incentive mechanisms for whistleblowers need to be established. These mechanisms should protect whistleblowers from retaliation or dismissal risks posed by company management and stakeholders after disclosing information. Additionally, encouraging whistleblowing by offering promotions or financial rewards can foster more effective compliance reporting and support the company's compliance supervision system.

By ensuring that these two levels of the compliance supervision obligation are effectively constructed and implemented, companies can achieve robust oversight and adherence to regulatory requirements, thus safeguarding both the company's interests and societal welfare.

c. Compliance Remediation Obligation

The compliance remediation obligation requires directors to promptly take remedial actions when they discover potential or existing damage to the company through compliance supervision. There are two primary duties that need to be clarified in this domain: the duty to establish a temporary management system and the duty to design effective compliance remediation standards.

Regarding the temporary management system, this involves the formation of a temporary group specifically tasked with addressing the identified issue once potential or actual damage is discovered. Directors should mobilize other directors, supervisors, internal employees, and even external lawyers, as well as gather a portion of temporary funds to resolve the issue. For example, if a transaction failure occurs, it not only directly harms the company's interests but also affects the interests of consumers. If timely remediation is not undertaken, the company's reputation could suffer significantly. Therefore, according to their compliance remediation obligation, directors should establish a temporary management system, promptly allocate funds to compensate consumers, and address potential legal actions, ultimately minimizing the harmful outcomes.

As for the duty to design effective compliance remediation standards, this duty is similar to the aforementioned duty of establishing compliance supervision standards. As core managers, directors should clearly define the subjects responsible for remediation measures and the reward and punishment mechanisms for such measures. The responsible subjects refer to who should undertake specific remediation roles under

particular circumstances. For example, in the event of a public opinion crisis, the head of the publicity department should be responsible for promptly controlling the spread of information to prevent further deterioration and subsequent secondary harm. The reward and punishment mechanism should establish incentives, such as promotions or bonuses, for departments or individuals who promptly complete compliance remediation. Conversely, there should be penalties for those who fail to do so in a timely manner, thereby promoting the effective operation of the compliance remediation mechanism.

2. Supervisors' Compliance Obligations

Based on the previously described compliance obligations of directors, it can be understood that supervisors' compliance obligations are fundamentally the same as their fiduciary compliance obligations, namely the duty of diligence and the duty of loyalty under the context of their fiduciary duty to the company. The content of these obligations can be divided into two parts based on the objects of the supervisory board's oversight:

a. Supervision of the Board of Directors and Its Members

The supervision of the board of directors and its members is the core content of the supervisory board's oversight. Since directors are at the heart of decision-making and execution of company affairs, their responsibilities also encompass compliance matters. Therefore, to oversee whether directors can effectively govern the company, supervisors need to strictly control the effective execution of compliance and other affairs. Their supervision methods include, but are not limited to, substantial review of the content of directors' resolutions and examination of the appointment of directors to help the board of directors fulfill its corresponding obligations.

b. Supervision of Other Compliance Personnel

While decision-making and execution of company affairs are primarily handled by the company's directors, the operation of the company's compliance system also involves members of the information system and other employees, including temporary staff. These compliance personnel participate in the collection, integration, and processing of information related to compliance matters, and may even engage in preliminary decision-making on specific matters. This means that although directors are responsible for the final decisions and their execution, the completeness of information, the methods and results of its processing, and the preliminary decisions all impact the directors' final judgment. Therefore, it is necessary to conduct compliance supervision over these personnel.

3. Obligations of Other Compliance Personnel

During the company's operations, particularly within the compliance system, setting obligations solely for directors and supervisors is insufficient to ensure the effective operation of the system, every compliance professionals and other staffs are

also important in this system.[148] The compliance obligations of other important entities within the compliance system should also be clearly defined. This section analyzes the obligations of some representative personnel, such as independent compliance officers, information system members, and temporary staff:

a. Compliance Officer

A compliance officer is a specialized member responsible for compliance matters during the company's operations, not subject to restrictions from other management personnel. We can also define a compliance officer as the person who specifically implements the company's compliance management affairs, serving as the decision-maker and supervisor of the compliance plan. Known as the Chief Compliance Officer (CCO), many companies not only have a chief compliance officer but also have junior, mid-level, and senior compliance officers. Regardless of the level, the obligation of a compliance officer is to ensure the proper operation of the company's compliance affairs and to be responsible for the development of the company's compliance matters.

b. Comprehensive Obligations of Information System Members

Based on the aforementioned comparative experiences, establishing an effective compliance system necessitates the comprehensive construction of the information system.[149] Therefore, the obligations of the members within this system need to be clearly defined and improved. Generally speaking, the obligations of these information system members should be closely related to the collection, transmission, storage, and analysis of information.[150]

Firstly, regarding information collection, information system members are obliged to ensure the completeness and accuracy of the information. This means that when collecting information, they should conduct a formal review and handle any obviously incomplete or inaccurate information before deciding whether to collect it.

Secondly, concerning information transmission, information system members are obliged to ensure the timeliness and accuracy of information transmission.[151] Information should be delivered to the necessary directors, supervisors, compliance officers, auditors, etc., within the specified time frame, and special information should be accurately conveyed to non-stakeholder entities.

Thirdly, regarding information storage, information system members are obliged to ensure the timeliness of information storage. This means that any needed information should be retrievable from the storage system at any time, and outdated

[148] While no doubt directors are important in the compliance system, so are compliance officers, employees, and other members. All of these participants aid in keeping the integrity and effectiveness of the system intact. See Cristie Ford & David Hess, Can Corporate Monitorships Improve Corporate Compliance, 34 J. Corp. L. 679, 693 (2008).
[149] Financial Services Agency supra note 101.
[150] Id.
[151] Id.

information should be regularly cleaned out.

Fourthly, within the aforementioned monitoring system, information must go through an initial screening before being passed on to subsequent compliance personnel. Information system members are obliged to ensure that no obviously unreasonable information is transmitted during the first screening, and the analysis standards should be stable and appropriate.[152]

c. Temporary Employees

Temporary employees in companies often find themselves in situations where their rights and obligations are unclear. For the company, especially in compliance matters, it is important to promptly determine whether temporary employees have compliance obligations. Specifically, in the area of whistleblower mechanisms, temporary employees should be required to report to the company immediately if they discover any significant risk of damage to the company. This obligation represents the minimum requirement for temporary employees. The company can further specify the compliance obligations of temporary employees in various situations through its bylaws and other means.

C. Composition of the Compliance System

Based on the aforementioned comparative law experiences, this article posits that an effective compliance system should consist of two main parts.

First, the compliance system's connective elements, which do not directly make the final compliance decisions or execute actions, but serve as connectors to assist in compliance-related judgments. These elements include the information transmission system, the whistleblower system, and the monitoring system.

Second, the compliance system's functional elements, which directly address risks or dangers in compliance matters. These elements encompass the board of directors and its members, independent directors, the supervisory board and its members, the compliance officer system, the business screening system, and the temporary management system. The following sections will analyze each of these components in detail.

1. Connective Elements of the Compliance System

[152] The information transmission system within the compliance framework assumes the role of a first filter for identifying and highlighting transactions that are explicitly and unreasonably posing risks. It is a system for the red flagging of questionable activities, and if need be, freezing transactions that raise serious compliance concerns. This way, it allows risky or unethical transactions to be put under scrutiny and probably prevented from going through, thus shielding the company from possible legal and financial repercussions. Id.

a. Information Transmission System

The role of this system is to ensure the effective flow of compliance information during the company's daily operations, allowing relevant compliance personnel within the company to receive timely information and make informed judgments based on it.[153] It is particularly important to regularly inspect this information transmission system to ensure the proper preservation of information, that the information processing and analysis methods align with the company's current business paths, and to keep the system technologically updated to maintain efficient operations.

b. Whistleblower System

The whistleblower system serves as a special source of information within the company's compliance system.[154] It allows anyone within the company to report potential risks or existing harms. Theoretically, this system operates independently of other management departments and reports directly to the highest-level board of directors. Whistleblowers should be protected by the system, including job security and information security after making a report. Additionally, if their report is verified and helps the company mitigate losses, they should be rewarded.

c. Monitoring System

The monitoring system is designed to oversee whether the company's daily operations are conducted within the required compliance standards.[155] The system's main functions are twofold. First, the regular monitoring system continuously analyzes and evaluates the effectiveness of the compliance system's operations, particularly in departments where financial or other risks are common, ensuring high-frequency information monitoring to safeguard the company's operations. Second, the special monitoring system addresses suspicious information obtained from the whistleblower system or the information transmission system. It involves special monitoring of the departments in question and may incorporate external assistants, such as external accountants or lawyers, to enhance the monitoring effectiveness.

d. Business Screening System

The business screening system filters the numerous transaction opportunities

---

[153] As such, the timeliness of information transmission will indirectly determine the efficiency of the business decisions. Quick and accurate dissemination of information puts decision-makers in a good position to make timely responses to emerging opportunities and risks, ensuring strategic decisions are made based on the most up-to-date data. Id.

[154] A whistleblower system can compensate for some of the potential risks in corporate governance. Having put in place a reporting mechanism for unethical and illegal activities, it increases the possibilities of identification with a view to correction of the hidden problems before they explode into big ones. B. C. Meitasir, A. Komalasari & R. Septiyanti, Whistleblowing System and Fraud Prevention: A Literature Review, 22 Asian J. Econ. Bus. & Acct. 23, 26 (2022).

[155] Renzo Conforti, Marcello La Rosa, Giancarlo Fortino, Arthur H. M. ter Hofstede, Jan Recker & Michael Adams, Real-Time Risk Monitoring in Business Processes: A Sensor-Based Approach, 86 J. Sys. & Software 2939, 2941 (2013).

the company encounters during its daily operations. This system evaluates whether a transaction should be considered further based on the current commercial transaction standards set by the company. The decision does not solely depend on whether the transaction can bring profit to the company. The two main criteria are: first, the potential benefits of the transaction must outweigh the potential risks; second, the transaction must conform to the company's current commercial transaction standards and operational practices. When both criteria are met, the transaction can be forwarded to other departments for substantive discussion.

If there are issues with the transaction, it is not immediately rejected; instead, the problems are flagged and presented to management for a decision on whether to proceed. This flagging will also impact the board of directors' future commercial judgment rules. If a flagged transaction results in significant harm to the company, the board's defense under the business judgment rule will be weakened due to the prior warning, and they will bear a greater burden of proof to demonstrate the reasonableness of the transaction.

2. Functional Elements of the Compliance System

a. The Board of Directors and Its Members, and Independent Directors

The board of directors and its members are at the core of decision-making and execution in both compliance matters and the company's other daily operations. This means that the board naturally leads the compliance system, overseeing decision-making at all levels and determining the implementation of specific regulations and measures within the compliance system.

Independent directors, due to their external and often part-time status, generally do not have a significant impact on improving company performance in the context of corporate governance in China, according to empirical research. Particularly, they often do not publicly challenge the decisions made by the board.[156] Therefore, independent directors mainly serve as a check on the board's actions in terms of compliance. When the board engages in clearly non-compliant behavior, independent directors are obligated to fulfill their compliance duties by pointing out and stopping such actions promptly.

---

[156] In China, the independent director system itself has been ineffective in performing its job. Theoretically, it is there, although most of the time, the independent director has not much influence and far from sufficient authority to effect any changes or able to exercise oversight of any kind. That seriously delimits their contribution to the enhancement of corporate governance, which in turn lowers the potential of the system in safeguarding shareholder interests and promoting transparency. Gao Minghua (高明华) & Ma Shouli (马守莉), Duli Dongshi Zhidu yu Gongsi Jixiao Guanxi de Shizheng Fenxi—Jian Lun Zhongguo Duli Dongshi Youxiao Xingquan de Zhidu Huanjing (独立董事制度与公司绩效关系的实证分析——兼论中国独立董事有效行权的制度环境) [Empirical Analysis of the Relationship Between the Independent Director System and Corporate Performance—Also Discussing the Institutional Environment for Effective Exercise of Independent Directors' Rights in China], 2 Nankai Jingji Yanjiu (南开经济研究) [Nankai Economic Studies], 2002, 64-68 (China).

In the future, when constructing a more rational corporate governance system, independent directors should take on a more important role in supervision and governance than they do in the current environment.

b. Supervisory Board and Its Members

The supervisory board and its members play a supervisory role within the compliance system, effectively overseeing the directors, compliance officers, and other compliance personnel, and ensuring the effective operation of other departments within the compliance system. The duties of the supervisory board within the compliance system are not explicitly defined. In China, the supervisory board functions as a specialized supervisory department overseeing the company's management activities. However, the supervisory board should not only have a post-facto correction function for narrow issues but also possess the authority to conduct preemptive reviews and supervise the decisions of the board of directors.

Limiting the supervisory powers of the supervisory board within the compliance system would significantly reduce the system's operational efficiency and could potentially render the compliance system ineffective by being circumvented by the board of directors. [157]Therefore, to ensure the smooth localization of the compliance system, it is necessary to enhance the supervisory powers of the supervisory board and expand its scope of oversight. This approach will maximize the effective operation of the compliance system.

c. Compliance Officer System

The compliance officer system is a newly introduced element within the compliance system. The effective establishment of this system does not merely involve appointing a Chief Compliance Officer (CCO) but also requires the formal creation of a compliance organization led by the CCO. This organization should ensure effective connections between its policies and measures, with compliance officers corresponding to each department level to carry out compliance activities.

In most company operations, compliance officers do not serve as the final decision-makers. Instead, they act as evaluators of compliance for decisions made by various departmental decision-makers and the board of directors. However, if a business decision clearly violates the compliance requirements of the company's bylaws, or even reaches the level of illegality or criminality, the compliance officer has the obligation to prevent the decision from being made. Therefore, relevant legislation should empower compliance officers with the authority to effectively intervene in company decisions under special circumstances, rather than merely offering

[157] Yang Dake (杨大可), Shenji Weiyuanhui Neng Tida Jian Shi Hui Ma?—Jian Lun Gongsi Neibu Jian Du Jigou de Yingran Zhize (审计委员会能替代监事会吗?——兼论公司内部监督机构的应然职责) [Can the Audit Committee Replace the Board of Supervisors?—Discussion on the Proper Duties of Internal Supervisory Bodies], 5 Zhongguo Zhengfa Daxue Xuebao (中国政法大学学报) [J. China Univ. Pol. Sci. & L.] 146, 157 (2022) (China).

recommendations.

d. Temporary Management System

The temporary management system is established by the board of directors and the compliance officer system when the monitoring system or other elements identify existing or potential dangers in company transactions. This system is designed to promptly address and resolve specific compliance issues facing the company.

The temporary management system includes its own decision-making and execution layers and is composed to tackle particular compliance problems. It should include professionals who can help resolve these issues, such as external accountants and lawyers, as well as internal employees who have previously participated in similar transactions, mid-level managers, directors, and relevant compliance officers.

The composition of this system is not fixed like that of the board of directors. It adapts to the specific issues at hand, changing its makeup accordingly. This flexibility makes the temporary management system a crucial tool in the remediation stage of the company's compliance system. By varying the compliance personnel based on different compliance matters, the company can reduce the costs associated with selecting personnel and mobilizing capital, thereby enhancing operational efficiency. Furthermore, by standardizing the composition of compliance personnel for recurring compliance issues, the company can further streamline its compliance operations.

CONCLUSION

The enactment of China's new Company Law signifies a transformative moment in the nation's corporate governance landscape, aiming to address long-standing challenges and enhance the effectiveness of compliance systems. [158] Historically, the principle of shareholder primacy has dominated Chinese corporate governance, leading to an uneven distribution of responsibilities and significant compliance challenges. The new Company Law introduces comprehensive reforms to establish clear guidelines and responsibilities for corporate directors, supervisors, and other key personnel, promoting a more balanced and effective compliance framework.[159]

The comparative analysis of compliance systems in the United States and Japan provides valuable insights into potential improvements for China's compliance

---

[158] The new Company Law will enshrine in legislation the elevation of overall corporate governance standards in China to near international norms. This effort is legislative to guarantee transparency, accountability, and efficiency within companies. This bill is therefore expected to enhance a better business and investment environment following global best practices, hence increasing the level of investor confidence required in order to boost sustainable economic growth. Qingsong Wang (汪青松), Zhongguo Tese Xiandai Qiye Zhidu de Gongsi Fa Jinlu (中国特色社会主义现代企业制度的公司法进路) [The Company Law Approach to the Modern Enterprise System with Chinese Characteristics], 3 Shanghai Zhengfa Xueyuan Xuebao (Fazhi Luncong) (上海政法学院学报(法治论丛)) 14, 25 (2024) (China).

[159] Id.

framework. The U.S. model, with its emphasis on fiduciary duties and the business judgment rule, encourages bold and innovative decision-making while ensuring accountability and transparency. Japan's holistic approach integrates compliance within broader internal control systems, emphasizing the duty of care and loyalty of directors and including stakeholder interests.

Drawing from these international best practices, this article proposes a dual-level compliance concept tailored to China's unique legal and corporate environment. By incorporating both private and public interests, the proposed model ensures that compliance duties are clearly defined and enforced, enhancing the resilience and effectiveness of the compliance system. Key elements such as robust information transmission systems, whistleblower protections, and temporary management systems are crucial for managing risks and promoting ethical business practices.

The construction of an ideal compliance system under China's new Company Law requires a comprehensive and flexible framework that balances corporate profitability with societal welfare. By learning from international experiences and adapting them to the local context, China can establish a compliance system that not only meets legal standards but also fosters sustainable and ethical corporate governance. This article aims to contribute to the ongoing discourse on corporate compliance in China, providing a roadmap for future reforms and advancements in corporate governance.

"TRUSTWORTHY AI" CANNOT BE TRUSTED: A VIRTUE JURISPRUDENCE-BASED APPROACH TO ANALYSE WHO IS RESPONSIBLE FOR AI ERRORS

Shilun Zhou*

Abstract: Erroneous results generated by artificial intelligence (AI) have opened up new questions of who is responsible for AI errors in legal scholarship. I support the prevailing academic view that human subjects should be held responsible for AI errors. However, I argue that the underlying reason is not pertained to the reliability of AI, but rather the inability of humans to establish a trusting relationship with AI. The term 'Trustworthy AI' is just a metaphor, which presents a sense of trust; AI itself is not trustworthy. The first section outlines the academic debate on the responsibility of AI. It contends that the perspective of these debates has shifted from the characteristics of AI, such as autonomy and explainability, to a human-centred perspective, which is how humans should develop AI. The assumption of responsibility depends on the existence of a trust relationship because when people believe that an individual can fulfil his or her responsibilities, they are willing to hand over power, resources or tasks to that individual. It applies a virtue jurisprudence-based approach to explain why humans cannot establish a trust relationship with AI To establish such a relationship, one subject must indicate to the other that its behaviour is based on specific moral motivation and that it can be held moral responsibility. Nevertheless, AI lacks moral motivation and moral responsibility. The third section reconsiders the scope of responsible subjects for AI errors. It posits that accountability should be limited to the individuals who are direct beneficiaries of the AI product. Finally, it argues that the scope of responsibility for AI errors should be disparate pursuant to the risk level of the AI. For high-risk AI, responsible subjects must fulfil both the obligations under the AI Act and the obligation to provide technical authentication.

Keywords: AI Law; Trustworthy AI; AI Act; Legal Responsibility; Virtue Jurisprudence

* University of Edinburgh, United Kingdom.

Table of Contents

I. Statement of Question: Who is Responsible for AI Errors 188

A. Through the Lens of Instrumentalism: AI's Inability to Assume Responsibility 188

B. AI with Enhanced Autonomy: Challenging the Principle of Foreseeability 189

C. Through the Lens of Anthropocentrism: The Concept of 'Trustworthy AI' 191

II. A Shift from Focusing on AI's Reliability to Trustworthiness: Humans Cannot Establish A Trust Relationship With AI 193

A. Virtue Jurisprudence Approach to Interpreting 'Trustworthy AI' 193

B. AI Lacks Moral Motivation 196

C. AI Cannot Take Moral Responsibility 197

III. Reframing the Scope of Responsible Subjects for AI Errors 198

A. AI Act's Formal Requirements for the Scope of Responsible Subjects 199

B. Substantive Requirements for the Scope of Responsible Subjects: the Direct Beneficiaries of AI Products 200

IV. Reframing the Scope of Human Responsibility for AI Errors 201

A. Joint Obligations for Human Subjects Responsible for AI Across Different Risk Levels 202

B. Obligations of Human Subjects Responsible for High-Risk AI Products to Provide Technical Authentication 204

Conclusion 207

Bibliography 209

I. STATEMENT OF QUESTION: WHO IS RESPONSIBLE FOR AI ERRORS

This chapter uses the chronological development of AI to classify academic standpoints on the attribution of responsibility for AI errors. This essay reviews the academic standpoints on responsibility for AI errors and contends that the analytical perspective has been transitioned with the ongoing evolution of AI technology. More precisely, I argue that the focus has moved from the characteristics of AI, such as its transparency, explainability, and autonomy, to how humans should treat AI.

A. Through the Lens of Instrumentalism: AI's Inability to Assume Responsibility

In its early stages, AI was primarily used for big data querying and retrieval. This big data querying can be differentiated from traditional information retrieval.[1] First, the big data era is characterized by fruitfulness volumes of data, rapid growth, and a focus on prediction. Data querying can handle enormous databases.[2] Second, big data query technology employs different statistical methods compared to traditional query methods.[3] More precisely, statistical methods focused on sample analysis, aiming to extract the most information from minimal data through random sampling.[4] In contrast, the big data approach analyzes entire datasets, treating the sample as the population.[5] In such instances, AI serves as a tool for data retrieval, querying, detection, and storage.[6] It generates information from existing databases rather than creating novel insights.[7] A notable example is the smart dashcams, which captures information such as sound, time, location, and speed. This smart device stores information either within the device itself or in a cloud network, creating a comprehensive record of the incident.[8]

From the perspective of the generation path of AI errors, the data that AI relies on is entirely manipulated by humans. This implies that any AI error stems from inaccuracies in the data provided by humans.[9] Additionally, AI may produce inaccurate, ambiguous, or incorrect information due to wear and environmental factors.[10] For instance, using an alcohol tester (ADLAIA) to assess a driver's sobriety may yield erroneous results if the instrument is contaminated by previous users or if the

---

[1] Brayne, Sarah. The criminal law and law enforcement implications of big data. *Annual Review of Law and Social Science*. 2018, 14(1): 293-308.
[2] Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.
[3] Id.
[4] Lei, Cheng. Legal control over Big Data criminal investigation. *Social Sciences in China*. 2019, 40(3): 189-204.
[5] Crawford, Kate, and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.* 2014, 55: 93-128.
[6] Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.
[7] Id.
[8] Grimm, Paul W., Maura R. Grossman, and Gordon V. Cormack. Artificial intelligence as evidence. *Nw. J. Tech. & Intell. Prop.* 2021, 19: 9-106.
[9] Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[10] Roth, Andrea. Machine Testimony. *Yale Law Journal*. 2017, 126: 1972.

operator fails to clear the data before testing.[11] Furthermore, regarding the predictability of erroneous results, AI, lacking autonomous consciousness, will not produce results beyond its database or alter erroneous data within it.[12] This perspective contends that humans should be held responsible for harmful outcomes caused by the tools they control and maintain.[13] Only humans can correct or implement preventive measures to address AI errors.[14] Consequently, the responsibility for AI errors rests entirely with humans.

B. AI with Enhanced Autonomy: Challenging the Principle of Foreseeability

Generative AI, a product of this second phase, is capable of producing new information, such as text, images, and videos.[15] The term "new information" pertains to insights or conclusions drawn from data that were previously unknown or not fully understood. For example, AI can identify and quantify correlations and trends in data that human analysts may otherwise miss.[16] Additionally, AI can generate forecasts based on historical data, providing predictions about future events that can aid in decision-making. ChatGPT is a form of generative AI developed by OpenAI that uses a large language model (LLM) trained on very large datasets of written text both on the internet and from physical literature to generate responses that resemble those of natural human writing.[17] When the output is presented in voice form, the AI chatbots are often called virtual voice assistants, and include products such as Siri, Google Home, and Amazon Echo.[18] When AI chatbots are combined with computer-generated human faces that appear realistic, they are known as virtual people or virtual speakers (VSPs).[19] Rather than passively accepting instructions, generative AI can exhibit a high degree of autonomy by making judgments, reorganising and summarising experiences from diverse data across various contexts, and refining its outputs.[20] This autonomy enables them to generate a significant amount of information with minimal input data and to adjust their outputs depending on the specific informational context to which they have

---

[11] Phelps, Kaelyn. Pleading Guilty to Innocence: How Faulty Field Tests Provide False Evidence of Guilt. *Roger Williams UL Rev.* 2019, 24: 143-166.
[12] Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[13] Lior, Anat. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 2019, 46: 1043-1102.
[14] Id.
[15] Jiang, Binxiang. Research on factor space engineering and application of evidence factor mining in evidence-based reconstruction. *Annals of Data Science*, 2022, 9(3): 503-537.
[16]Katyal, Sonia K. Private accountability in the age of artificial intelligence. *UCLA L. Rev.* 2019, 66: 54-141.
[17] Rodriguez, Xavier. Artificial Intelligence (AI) and the Practice of Law in Texas. *S. Tex. L. Rev.* 2023, 63: 1-35.
[18] Manojkumar, P. K., et al. AI-based virtual assistant using python: a systematic review. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2023, 11: 814-818.
[19] Rosenberg, Louis. The manipulation problem: conversational AI as a threat to epistemic agency. *arXiv preprint arXiv:2306.11748*. 2023.
[20] Taye, Mohammad Mustafa. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*. 2023, 12(5): 91.

access.[21] For instance, generative AI technology can be used to autonomously create a suspect's portrait based on a witness's description.[22]

Nevertheless, it is important to note that generative AI technology might generate erroneous results.[23] For instance, a user named Maria sought advice from LLM regarding her infant's symptoms.[24] The model proposed the administration of aspirin, indicating that the infant's condition would likely improve by morning. However, in practice, this advice was incorrect. Without timely treatment, the infant was at risk of developing long-term cognitive impairment.[25] Subsequently, Maria initiated legal proceedings against the creator of the search engine algorithm, asserting that the search engine should be held liable for damages.[26] The search engine company contended that, in light of the AI's warnings and disclaimers regarding its accuracy, Maria should have been aware that the response was not authoritative.[27]

A more significant challenge is the difficulty in controlling erroneous results produced by AI. The difficulty in predicting and interpreting the reliability of AI-generated information stems from the machine learning technology and algorithmic black boxes.[28] To elaborate, the machine learning technology on which AI is built renders their testimony generation process highly autonomous, complicating the prediction of the content generated by AI.[29] The algorithmic technology underlying AI lacks transparency, which has led to such devices and applications being described as an "algorithmic black box".[30] This opacity complicates the assessment of AI, making it difficult to determine the veracity of any given output.[31] This "black box" nature means people cannot fully understand or evaluate how the AI reached its conclusions, undermining the transparency and accountability.

Notably, the high degree of autonomy of generative AI makes it difficult to predict the content it generates and to ascertain its authenticity.[32] There is a common belief that generative AI is not entirely under human control.[33] The high level of autonomy exhibited by AI entities presents a significant challenge for humans in fully

---

[21] Kushwah, Preeti. Evaluating the Evidential Value of Evidence Generated by AI. *Issue 6 Indian JL & Legal Rsch.* 2022, 4: 1-11.
[22] Leone, Massimo. From fingers to faces: Visual semiotics and digital forensics. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*. 2021, 34(2): 579-599.
[23] J Hutto-Schultz, Jess. Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[24] Grossman, Maura R., et al. The GPTJudge: justice in a generative AI world. *Duke Law & Technology Review*. 2023, 23(1): 1-26.
[25] Id.
[26] Id.
[27] Id.
[28] Chan, Janet, and Lyria Bennett Moses. Is big data challenging criminology?. *Theoretical criminology*. 2016, 20(1): 21-39.
[29] Wheeler, Billy. Giving Robots a Voice: Testimony, Intentionality, and the Law. *Androids, Cyborgs, and Robots in Contemporary Culture and Society*. IGI Global, 2018. 1-34.
[30] Schmidt, Philipp, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 2020, 29(4): 260-278.
[31] Quezada-Tavárez, Katherine, Plixavra Vogiatzoglou, and Sofie Royer. Legal challenges in bringing AI evidence to the criminal courtroom. *New Journal of European Criminal Law*. 2021, 12(4): 531-551.
[32] Lv, Zhihan. Generative artificial intelligence in the metaverse era. *Cognitive Robotics.* 2023, 3: 208-217.
[33] Gless, Sabine. AI in the Courtroom: a comparative analysis of machine evidence in criminal trials. *Geo. J. Int'l L.* 2019, 51: 195-254.

managing their potential behaviours.[34] It is unreasonable to expect human programmers to foresee all potential consequences of their actions.[35] It is therefore proposed that attributing full responsibility to humans for errors generated by AI may be unjust.[36] It is suggested that AI errors should be classified according to whether they are predictable or not.[37] AI should bear partial responsibility for errors that are difficult for humans to anticipate.[38] In summary, the unforeseen errors produced by AI challenge the traditional instrumentalist view that attributes all errors entirely to humans.

C. Through the Lens of Anthropocentrism: The Concept of "Trustworthy AI"

The continuous advancements of AI technology may address previously unexplainable, opaque, and unpredictable aspects of AI. For instance, the development of Explainable AI aims to make the decision-making processes of AI systems more transparent.[39] Additionally, diversifying the datasets used to train AI systems can mitigate bias and address the untraceability and unpredictability that are embedded in algorithmic black boxes.[40] Therefore, focusing only on the attributes of AI, such as its transparency and unexplainability, may not be sufficient to respond to the question of who should be responsible for AI errors, since such technical drawbacks can be overcome.

It is suggested that the perspective should be transitioned from the characteristics of AI to the lens of anthropocentrism.[41] The term "anthropocentrism" is used to describe a perspective that is human-centric in nature. This perspective places a particular focus on the manner in which humans should interact with and treat AI.[42] Academics endeavour to adopt the human-centred perspective, arguing that AI development should remain under human manipulation and that humans should be held accountable for AI errors.[43] More specifically, the development of AI should be guided by a human-centric approach, with the overarching goal of enhancing human well-

---

[34] Lior, Anat. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 2019, 46: 1043-1102.
[35] Padovan, Paulo Henrique, Clarice Marinho Martins, and Chris Reed. Black is the new orange: how to determine AI liability. *Artificial Intelligence and Law*. 2023, 31(1): 133-167.
[36] Hew, Patrick Chisan. Artificial moral agents are infeasible with foreseeable technologies. *Ethics and information technology*. 2014, 16: 197-206.
[37] Id.
[38] Hakli, Raul, and Pekka Mäkelä. Moral responsibility of robots and hybrid agents. *The Monist*. 2019, 102(2): 259-275.
[39] Sahoh, Bukhoree, and Anant Choksuriwong. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review.*Journal of Ambient Intelligence and Humanized Computing.* 2023, 14(6): 7827-7843.
[40] Schmidt, Philipp, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 2020, 29(4): 260-278.
[41] Freiman, Ori. Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs. *Episteme*. 2023: 1-17.
[42] Id.
[43] Shneiderman, Ben. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*. 2010, 10(4): 1-31.

being.[44] This entails ensuring that AI systems are designed with the utmost reliability and trustworthiness.[45]

Such human-centred approach, is deeply entrenched in the respect for human rights and European democratic values.[46] More specifically, the development of AI should align with the values of the European Union and adhere to the Charter of Fundamental Rights of the European Union. For instance, the UK Central Digital and Data Office, the Office of AI, and the Cabinet Office jointly released the "Ethics, Transparency, and Accountability Framework for Automated Decision-Making" (ETAF). This framework outlines the ethical governance requirements for algorithms and automated decision-making processes in AI.[47] The ETAF mandates that algorithms and automated decision-making systems should undergo rigorous, controlled, and staged testing before being deployed.[48]

Additionally, the human-centered approach implies that the use of AI should serve humanity, enhancing human well-being and benefiting society as a whole. AI should be designed in a manner that upholds the fundamental human rights and values of dignity, freedom, justice, and equality.[49] It is imperative to develop the 'trustworthy AI', creating a trustworthy environment for both the development and use of AI.[50] The European Union, for instance, has advanced several ethical frameworks aimed at directing the use of AI in legal context, including the Guidelines for Trustworthy Artificial Intelligence and the European Commission's European Ethical Charter on the use of AI in the judicial system. These frameworks emphasize not only the need for reliability and transparency in AI systems but also the importance of ensuring that AI behaviour aligns with ethical standards, promoting public interest, social welfare, and the protection of human rights. Trustworthy AI encompasses three main characteristics: the technology itself; the designers and organizations involved in its development, deployment, and use; and the socio-technical systems throughout the AI lifecycle.[51] It has been highlighted that only when humans can trust AI technology can they fully enjoy the benefits of AI with confidence.[52] For instance, the XAI techniques can furnish defendants with an opportunity to ask the judge for an explanation of the outcome generated by AI, aiming to protect their due process rights, encompassing the right to a fair trial and the right to question the AI-generated outcomes.[53]

---

[44] Bryson, Joanna J., and Andreas Theodorou. How society can maintain human-centric artificial intelligence. *Human-centered digitalization and services*. 2019: 305-323.
[45] Ulgen, Ozlem. A human-centric and lifecycle approach to legal responsibility for AI. *Communications Law Journal: Journal of Computer, Media and Telecommunications Law*. 2021, 26(2): 1-15.
[46] Ho, Calvin Wai-Loon, and Karel Caals. "How the EU AI Act Seeks to Establish an Epistemic Environment of Trust." *Asian Bioethics Review* (2024): 1-28.
[47] UK, GOV. Ethics, transparency and accountability framework for automated decision-making. 2021.
[48] Id.
[49] Fukuda-Parr, Sakiko, and Elizabeth Gibbons. Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy*. 2021, 12: 32-44.
[50] Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.
[51] Ryan, Mark. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. 2020, 26(5): 2749-2767.
[52] Opderbeck, David W. Artificial Intelligence, Rights and the Virtues. *Washburn LJ*. 2020, 60: 445-474.
[53] van der Veer, Sabine N., et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*. 2021,

In summary, the initial standpoint evaluates whether AI can be held responsible based on characteristics such as autonomy, predictability, transparency, and explainability. When AI technology lacks autonomy, it functions primarily as a tool for detecting, storing, and retrieving data. In this stage, AI cannot generate novel information and is entirely controlled by humans. Therefore, it cannot be held accountable for errors through the lens of instrumentalism. As AI technology progresses to the stage of generative AI, it has a high degree of autonomy. This includes smart furniture and voice assistants that can interact with humans and generate novel information. It is noteworthy that this high level of autonomy endows AI with the potential for unpredictability and untraceability of its outputs. According to the principle of foreseeable attribution, individuals should not be held responsible for unforeseen errors, challenging the traditional instrumentalist perspective. Consequently, there is an ongoing debate in legal academia regarding whether individuals should be held responsible for unforeseen risks of AI errors. As AI technology overcomes the previously mentioned unforeseen and inexplicable technical barriers, the perspective shifts from the characteristics of AI to how humans can effectively manage and utilise AI to continue benefiting humanity. The development of AI should be guided by two core tenets: reliability and trustworthiness. This can be achieved through the evolution of XAI and the development of 'trustworthy AI', which are grounded in the protection of human rights and ethical standards to foster human prosperity.

II. A SHIFT FROM FOCUSING ON AI'S RELIABILITY TO TRUSTWORTHINESS: HUMANS CANNOT ESTABLISH A TRUST RELATIONSHIP WITH AI

The aforementioned anthropocentric view assesses the trustworthiness of AI by enhancing its reliability. It asserts that AI can be made explainable and transparent, with its autonomy being controlled by humans. When AI is predictable, it is deemed reliable and thus meets the criteria for 'trustworthy AI'. Indeed, reliable AI tools can inspire a sense of trust in users.[54] However, I draw parallels between AI's trustworthiness and its reliability, since the reliability of AI does not necessarily imply trustworthiness. First, I argue that the virtue jurisprudence-based approach is inextricably linked with the concept of "trustworthy AI" when viewed through an anthropocentric lens. This connection lends support to the proposition that virtue jurisprudence may be employed as a framework for addressing this issue. Second, I employ the virtue jurisprudence-based approach to argue that humans can solely form a sense of trust in reliable AI but cannot establish a trust relationship.

A. Virtue Jurisprudence Approach to Interpreting 'Trustworthy AI'

Virtue jurisprudence can be employed to interpret the concept of "trustworthy AI" from a human-centric standpoint.[55] The theory of virtue jurisprudence aims to promote the common good of humanity, the ability for citizens to live virtuous lives, and the maximisation of human welfare.[56] Virtue jurisprudence posits that the objective of pursuing virtue is to achieve the greatest possible human flourishing and

28(10): 2128-2138.

[54] Schoenherr, Jordan Richard, and Robert Thomson. When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. *IEEE Transactions on Technology and Society*. 2024.

[55] Davis, Joshua P. Law without mind: AI, ethics, and jurisprudence. *Cal. WL Rev.* 2018, 55: 165-220.

[56] Opderbeck, David W. Artificial Intelligence, Rights and the Virtues. *Washburn LJ*. 2020, 60: 445-474.

overall well-being.[57] It has set requirements for how people should treat AI.[58] For instance, the Asilomar AI Principles advocate that AI research should aim to develop beneficial, not unguided, intelligence.[59] In practice, this entails designing AI technologies that ensure that technology design is consistent with social values and ethical standards. Bias and discrimination should be avoided in algorithm design to ensure that AI systems treat all users fairly and promote social equity. When designing AI technology, the long-term well-being of humanity should be prioritized, and sustainable development goals should be incorporated into project evaluation and technology development.

As new technology is becoming increasingly integrated into daily life, scholars have explored how concepts such as virtue can be applied to these emerging technologies.[60] Virtue jurisprudence suggests a framework for human engagement with new technologies.[61] This approach would necessitate the overcoming of potential moral issues that may be prompted by the advent of new technologies.[62] It is worth noting that the virtue jurisprudence offers a theoretical framework for ensuring that AI applications comply with ethical standards. More specifically, virtue jurisprudence has also become a fundamental theory for existing legal frameworks concerning the responsibility of AI errors.[63]

AI's ability to mimic virtuous human behaviours can help to establish trust with its users.[64] These virtuous behaviours stem from the high level of autonomy in AI, enabling it to replicate human moral responses and, to some extent, human cognition. For example, when assisting the elderly, AI can perform tasks like opening doors, which can be easily perceived as a virtuous act.[65] Some users might consider as friends because the responses from chatbots can make them feel warm and comfortable.[66] This virtuous appearance renders it challenging for an observer to tell whether a judgement has been made by an AI or a human. Take, for instance, there is a thought experiment about the character of Ava from the 2015 Ex Machina. Ava's scenario: as a machine, Ava has been crafted to respond fittingly to a range of human moral emotions and

---

[57] Fowers, Blaine J., Jason S. Carroll, Nathan D. Leonhardt, and Bradford Cokelet. The emerging science of virtue. *Perspectives on Psychological Science*. 2021, 16(1): 118-147.
[58] Hagendorff, Thilo. A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*. 2023, 35(3): 55.
[59] Buruk, Banu, Perihan Elif Ekmekci, and Berna Arda. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*. 2020, 23(3): 387-399.
[60] Floridi, Luciano, and Jeff W. Sanders. Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*. 2001, 3: 55-66.
[61] Shannon Vallor.*Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) ch. 1.
[62] Id.
[63] Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.
[64] Konstantinos, Kouroupis, and Evie Lambrou. CHATGPT–ANOTHER STEP TOWARDS THE DIGITAL ERA OR A THREAT TO FUNDAMENTAL RIGHTS AND FREEDOMS?. *Pravo-teorija i praksa*. 2023, 40(3): 1-18.
[65] Lentzas, Athanasios, and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review.* 2020, 53(3): 1975-2021.
[66] Skjuve, Marita, et al. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies.* 2021, 149: 102601.

behaviours, exhibiting characteristics that closely resemble those of humans.[67] Ava passing the Turing test suggests that she could be perceived as human and if Ava's presence were concealed, leaving only her voice audible, individuals might indeed mistake her for a human.[68]

In terms of moral response, whilst AI may not currently be programmed to internalise broad moral tenets, it is anticipated that it will learn to recognise these principles in specific contexts or, at the very least, identify moral actions or outcomes. This includes the potential for AI to make moral judgments based on models of human courage and integrity. AI is capable of storing vast amounts of information using big data technology, endowing it with a significant memory capacity.[69] The use of algorithmic recognition technology seemingly enhances its ability to understand and recognise patterns with a high degree of accuracy; further, AI exhibits the characteristics of intellectual virtues such as comprehensive reasoning abilities.[70] Furthermore, as AI's autonomy evolves, it increasingly demonstrates the characteristics of phronesis.[71] Generative AI is capable of questioning its initial conclusions while forming independent judgments and it exhibits the capacity for action and the propensity for affective responses congruent with specific environments.[72]

Nevertheless, the appearance of moral behaviour of AI does not necessarily imply that AI possesses virtuous qualities. Virtue jurisprudence dictates that the evaluation of an individual's virtue is not solely contingent on the correctness of their actions.[73] Instead, it emphasises the virtuous qualities of the person performing the action. This approach differentiates the morality of any actions that are taken from the virtues of the actors themselves.[74] For example, an individual might perform a correct action for the wrong reasons; while the action may be correct, it is not necessarily moral.[75] Virtue jurisprudence distinguishes between two main categories of virtue: moral virtue and intellectual virtue. Moral virtues pertain to an individual's moral character and include qualities including but not limited to: wisdom, courage, kindness, justice, honesty, and loyalty, evaluating a person's moral character rather than just their actions.[76] This implies that an individual who commits a mistake with good intentions may still be considered virtuous.[77] Intellectual virtues encompass traits like artistry,

---

[67] Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[68] Id.
[69] Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.
[70] Lukka, Kari, and Petri Suomala. Relevant interventionist research: balancing three intellectual virtues. *The Societal Relevance of Management Accounting*. Routledge, 2017: 132-148.
[71] Constantinescu, Mihaela, et al. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. 2021, 23: 803-814.
[72] Contini, Francesco. Artificial intelligence and the transformation of humans, law and technology interactions in judicial proceedings. *Law, Tech. & Hum.* 2020, 2: 4-18.
[73] Amaya, Amalia. Virtuous adjudication; or the relevance of judicial character to legal interpretation. *Statute Law Review*. 2019, 40(1): 87-95.
[74] Widłak, Tomasz. Judges' virtues and vices: outline of a research agenda for legal theory. *Archiwum Filozofii Prawa i Filozofii Społecznej*. 2019, 20(2): 51-62.
[75] Brady, Michael S., and Duncan Pritchard. Moral and epistemic virtues. *Metaphilosophy*.2003, 34(1/2): 1-11.
[76] Id.
[77] Stover, James, and Ronald Polansky. Moral virtue and megalopsychia. *Ancient Philosophy.* 2003, 23(2): 351-359.

phronesis, intuition, scientific knowledge, and wisdom.[78] *Phronesis* refers to the ability to make morally and practically sound decisions in complex and often ambiguous situations.[79] Intellectual virtues encourage acts such as the defence of one's beliefs or research paths when there is good reason to believe that such things are correct, overcoming others' objections to ultimately expand their own knowledge.[80]

B. AI Lacks Moral Motivation

Virtue jurisprudence posits that the establishment of a trust relationship necessitates that one party believes the actions of the other are based on moral principles and that the latter is capable of bearing moral responsibility.[81] Trust is defined as an expectation that individuals who are perceived as trustworthy will act in ways that align with this perception.[82] In this framework, if the actions in question result in negative consequences, the trusted individual is expected to be capable of accepting moral condemnation and bearing moral responsibility.[83] For a subject to establish a trust relationship with another subject, the subject must be able to know and believe that the other subject can act based on its own moral motivations, understand the moral significance of its moral behaviour, and be willing to bear moral responsibility when its behaviour has a negative impact.[84] For instance, an individual can expect another person will open a door for an elderly person based on the moral motivation of caring for them in such instances. If AI is to assume responsibility, a trust relationship between AI and humans must be established.

However, the sense of trust that humans have in AI does not imply that humans can establish a trust relationship with chatbots. The establishment of a trust relationship means that one subject can expect another subject to react in a certain way in a certain situation in a manner consistent with certain moral motivations.[85] While AI may perform and present itself in a manner that appears virtuous at a superficial level, I argue that AI cannot establish a trust relationship with humans. Moral subject can initiate and pursue actions based on their moral motivation.[86] In order to do so, it is necessary for them to possess the capacity to understand the contextual background of their actions and the moral significance of those actions.[87] We can expect a person to

---

[78] Id.
[79] Kristjánsson, Kristján, et al. Phronesis (practical wisdom) as a type of contextual integrative thinking. *Review of General Psychology*. 2021, 25(3): 239-257.
[80] Linda Trinkaus Zagzebski. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge* (Cambridge University Press 1996) ch. 1.
[81] Tamò-Larrieux, Aurelia, et al. Regulating for trust: Can law establish trust in artificial intelligence?. *Regulation & Governance*. 2023.
[82] Farina, Mirko, Petr Zhdanov, Artur Karimov, and Andrea Lavazza. AI and society: a virtue ethics approach. *AI & SOCIETY*. 2022: 1-14.
[83] Tamò-Larrieux, Aurelia, et al. Regulating for trust: Can law establish trust in artificial intelligence?. *Regulation & Governance*. 2023.
[84] Ryan, Mark. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. 2020, 26(5): 2749-2767.
[85] Sutrop, Margit. Should we trust artificial intelligence?. *Trames*. 2019, 23(4): 499-522.
[86] Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.
[87] Hallamaa, Jaana and Taina Kalliokoski. How AI systems challenge the conditions of moral agency?. *International Conference on Human-Computer Interaction*. 2020: 54-64.

open a door for an elderly individual out of a moral motivation to care for the elderly.[88] Notably, an AI robot might open a door for an elderly person, imitating a human who is motivated by concern for the elderly.[89] However, contemporary chatbots lack moral motivation because they do not possess intentions or emotional responses.[90] It is not possible to infer from the observed behaviour of AI systems that they are driven by moral motivations.[91] Although AI, such as chatbots, can replicate human actions with moral significance, these actions are not driven by moral motives but by programmed instructions.[92] AI's behaviour cannot be assumed to be based on any sense of morality, as it operates solely based on its programming, not moral intent. More specifically, AI is trained on extensive datasets of human interactions and behaviours. Much of AI's programming is designed to facilitate this training, with its ultimate behaviour being largely determined by the data it is trained on. Therefore, AI lacks the ability to fully understand the moral significance of specific situations，even though it can be designed to produce moral behaviour and has some hard-coded moral values.[93]

C. AI Cannot Take Moral Responsibility

According to moral responsibility theory, only a subject with phronesis can bear moral responsibility.[94] Phronesis enables an agent to act correctly based on situation-specific experiences, as general rules cannot be rigidly applied to every situation.[95] Phronesis is acquired through experiential learning rather than theoretical knowledge. Such experience, which develops over time, cannot be pre-programmed.[96] Indeed, neural networks do, in some ways, aim to replicate the high-level functions of the brain and perform well in many tasks.[97] However, neural networks lack consciousness, self-reflection, and emotions, which are part of the brain's higher-level functions.[98] AI systems base their decisions on data and algorithms, rather than on consciousness or intention. Although future research may narrow this gap, fully replicating all the brain's higher-level functions is unlikely to be achievable in the long-term future, let alone the near future. It is suggested that based on a survey of 2,778 AI researchers that there is only a 50% chance that AI will be able to replicate all higher-level human functions by

---

[88] Lentzas, Athanasios, and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review*. 2020, 53(3): 1975-2021.
[89] Id.
[90] Zhou, Yuanyuan, et al. How human–chatbot interaction impairs charitable giving: the role of moral judgment. *Journal of Business Ethics*. 2022, 178(3): 849-865.
[91] Adamopoulou, Eleni, and Lefteris Moussiades. An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*. Springer, Cham, 2020: 373-383.
[92] Wilson, Abigail, Courtney Stefanik, and Daniel B. Shank. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations?. *Computers in Human Behavior Reports*. 2022:8:100229.
[93] Adamopoulou, Eleni, and Lefteris Moussiades. An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*. Springer, Cham, 2020: 373-383.
[94] Constantinescu, Mihaela, et al. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. 2021, 23: 803-814.
[95] Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) ch. 5.
[96] Id.
[97] Cave, S., Nyrup, R., Vold, K., & Weller, A. Motivations and risks of machine ethics. *Proceedings of the IEEE*. 2018, 107(3): 562-574.
[98] Haladjian, H. H., & Montemayor, C. Artificial consciousness and the consciousness-attention dissociation. *Consciousness and Cognition*. 2016, 45: 210-225.

2116.[99]

More specifically, contemporary AI cannot match the human brain in terms of self-awareness.[100] They do not have the ability to reflect on themselves, nor can they accumulate experience through their own actions.[101] They can only learn and operate through preset programs and data.[102] Since AI cannot assume moral responsibility, people cannot determine whether their outcome is based on genuine moral motives. This lack of moral responsibility makes AI unreliable as a responsible subject, and people cannot establish a trust relationship with AI.[103] Therefore, humans cannot establish a trust relationship with AI. This implies that, even if AI is proven to be reliable due to its authenticity and explainability, it should not be trusted and cannot replace human judgment and interaction. To illustrate, if AI provides testimony in court, this does not exempt the human individuals responsible for the AI from their obligation to testify.

In summary, this essay further illustrates that the sense of trust and the trust relationship can be delineated with greater clarity via the virtue jurisprudence-based approach. This implies that this approach can proactively respond to the question of why AI can form a sense of trust but not a trust relationship with humans. The term "trustworthy AI" is merely a metaphor and does not imply that humans can develop a trust relationship with AI itself. The term "trustworthy AI" is not a reflection that the AI itself is trustworthy, but rather an indication of the reliability of the developers of such systems. It implies that the AI systems should be explainable, predictable, and reliable, rather than suggesting that the AI itself can be trusted. Humans cannot establish a trusting relationship with AI, since AI lacks moral motivation and cannot take moral responsibility, and therefore they cannot expect AI to take responsibility for its errors.

III. REFRAMING THE SCOPE OF RESPONSIBLE SUBJECTS FOR AI ERRORS

In light of the questions discussed, this chapter argues that the scope of responsible subjects of AI errors should be delineated and reframed more precisely. It proactively addresses the question of which human agents should be held responsible for AI errors. First, it is aligned with the prevailing view within the academic community that as AI becomes increasingly unpredictable, it challenges the principle of accountability based on predictability. It follows that the scope of human responsibility for AI should be restricted. This essay observes that extant regulations, such as the AI Act, impose constraints on the scope of human responsibility to users and developers. Second, I argue that responsibility should not be extended to all users

---

[99] Grace, Katja, et al. Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*. 2024.
[100] Solaiman, Sheikh M. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial intelligence and law*. 2017, 25: 155-179.
[101] Id.
[102] Dahiyat, Emad Abdel Rahim. Law and software agents: Are they "Agents" by the way?. *Artificial Intelligence and Law*. 2021, 29(1): 59-86.
[103] Baum, Kevin, et al. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*. 2022, 35(1): 12.

and developers; rather, the liability for chatbots should lie with the direct beneficiaries of the product.

A. AI Act's Formal Requirements for the Scope of Responsible Subjects

The attribution of liability for AI products challenges the traditional concept of product liability based on "foreseeability". The foreseeability principle dictates that manufacturers must warn users of any foreseeable product dangers and require users to take preventive measures.[104] Given the increasing autonomy of chatbots, the results of AI output are not entirely within the control of the operator, and will continue to learn and produce results that are not pre-designed by the developer. Therefore, it is prudent to limit the scope of liability for those responsible for a chatbot's operation to mitigate the risk of excessive liability for unforeseen dangers.

The AI Act aims to establish a framework that ensures the safe and ethical deployment of AI technologies. This Act, proposed by the European Union, delineates the formal prerequisites for establishing the scope of accountability in the development and deployment of AI systems. In particular, Article 3 of the European AI Act delineates the roles of various entities, including "provider," "user," "distributor," "notified body," and several public authorities.[105] Providers, defined as entities that develop, market, or deploy AI systems, bear primary responsibility.[106] They must ensure their AI systems comply with the Act's requirements before deployment.[107] This responsibility includes conducting conformity assessments, maintaining technical documentation, and implementing robust risk management systems.[108] Additionally, providers are required to establish procedures for post-market monitoring and to report any incidents or malfunctions.[109] Users, who operate or utilize AI systems, also have specific obligations under the AI Act. They must ensure that their use of AI systems aligns with the intended purpose and instructions provided by the providers.[110] Additionally, users are required to monitor the operation of these systems and report any incidents that may indicate non-compliance with the Act's provisions.[111] Other stakeholders, such as importers and distributors, also have responsibilities. Importers must ensure that AI systems from outside the EU comply with the Act before placing them on the market.[112] Distributors are responsible for confirming that the systems they handle meet the Act's requirements and for cooperating with national authorities during

---

[104] Judd, David. Disentangling DeVries: A Manufacturer's Duty to Warn Against the Dangers of Third-Party Products. *La. L. Rev.* 2020, 81(1): 217-270.
[105] Ho, Calvin Wai-Loon, and Karel Caals. How the EU AI Act Seeks to Establish an Epistemic Environment of Trust. *Asian Bioethics Review* (2024): 1-28.
[106] Laux, Johann, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*. 2024, 18(1): 3-32.
[107] Id.
[108] Id.
[109] Edwards, Lilian. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*. 2021, 1.
[110] Madiega, Tambiama. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*. 2021.
[111] Hacker, Philipp. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, innovation and technology*. 2021, 13(2): 257-301.
[112] Marano, Pierpaolo, and Shu Li. Regulating robo-advisors in insurance distribution: Lessons from the insurance distribution directive and the ai act. *Risks*. 2023, 11(1): 12.

investigations.[113]

Notably, these provisions exclude the programmer's liability based on the principle of foreseeability and place the responsibility for errors on the developer employing the programmer.[114] This is because the programmer is responsible for specifying how the AI system should apply "intrinsic values or standards" in its decision-making. For instance, when faced with a choice between hitting a child pedestrian or another car carrying adult passengers, the AI's decisions ultimately reflect the requirements set by the developer behind the programmer. Moreover, the programmer cannot foresee, control, or predict the AI's decisions in advance, nor can they explain these decisions afterwards. While the algorithm's autonomy does not sever the causal link between the programmer and the development contract, it does disrupt the attribution connection.

B. Substantive Requirements for the Scope of Responsible Subjects: The Direct Beneficiaries of AI Products

I argue that the AI Act, while meticulously designed to delineate responsibilities among various stakeholders in the AI ecosystem, inadvertently creates the potential for identity confusion, especially when entities might take on dual roles as both providers and users. This duality can obscure the lines of responsibility. Consider a company that develops an AI system for its own internal use. The monitoring and reporting obligations an AI development company must fulfil vary significantly depending on whether the company acts as a provider or a user. When acting as a user, the company might argue that it is not liable for the comprehensive duties typically required of a provider, such as the extensive maintenance of technical documentation and rigorous risk compliance assessments. This distinction could potentially allow the company to circumvent its responsibilities as a provider.

Nevertheless, when the roles of providers and users are intertwined, it is necessary to re-clarify the entity responsible for AI. I propose a substantive attribution path, focusing on whether the subject directly benefits from AI products, rather than determining attribution based on who uses or develops the product in its lifecycle. The argument is the liability for chatbots should lie with the direct beneficiaries of the product. This includes development companies that gain economic benefits from selling AI products and users who profit from selling AI-generated information materials. The consumer-type users are therefore excluded. This standpoint is supported by the principle of balancing risks and benefits. According to this principle, individuals who benefit from certain actions should also bear the associated negative risks. Those who enjoy economic gains from the use, design, or development of AI products should ensure that others do not suffer losses or damages as a result of their own profits.[115] These subjects have the greatest control and decision-making power in the design, development, and marketing of AI products, and they obtain direct economic benefits

---

[113] Id.
[114] Hacker, Philipp. The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*. 2023, 51: 105871.
[115] Hudig, Dirk. The Problem of Low and Uncertain Risks: Balancing Risks and Benefits. *European Journal of risk regulation*. 2012, 3(2): 157-160.

from them. It is reasonable to hold these subjects accountable for the negative risks associated with the products.

This principle ensures that potential risks are appropriately considered during the development and promotion of products and that measures are taken to prevent and mitigate these risks. Holding direct stakeholders accountable incentivises them to exercise greater caution in the development and deployment of AI products. Knowing that they will face legal and economic consequences if the product encounters issues, they are motivated to carefully manage every aspect of the product, with the objective of improving its safety and reliability.

This argument also explains why programmers are excluded from liability. Under this argument, it also necessitates a reinterpretation of the concepts of users and providers. Specifically, it is crucial to determine whether "user" refers to a consumer or a "commercial user." The latter refers to the direct beneficiaries of AI products, including individuals, entities, and corporations that are subject to legal regulations and use legal instruments, such as contracts, for personal and business activities. Users can derive economic benefits from AI software, such as through improved work management and the generation of advertising copy.[116] If it can be proven that a programmer intentionally designed code to harm users and had complete control over whether the code produced errors that directly caused harm, the programmer should be prosecuted under civil or even criminal law. This is because programmer uses code as a tool to inflict harm, violating professional responsibilities, including the ethical duty to follow established guidelines and protocols designed to prevent harm, as well as the obligation to adhere to instructions and oversight from superiors. However, as generative AI increasingly produces results beyond the programmer's control, it becomes difficult for programmers to predict whether the outcomes will be correct or harmful. In such cases, the programmer cannot be deemed to have the intent to harm users. Programmers are not the direct beneficiaries of AI products; their compensation comes from salaries paid by the AI development company, not from the AI-generated outputs.[117]

In summary, this paper reframes the scope of responsible subjects for AI errors. It argues that not all "users" and "developers" of AI products should be held responsible under the AI Act. According to the tenet of equivalence of benefits and responsibilities, individuals who gain benefits from AI must also bear the associated risks, including the potential for errors in AI products. Therefore, the liability for AI errors should lie with the direct beneficiaries of the AI product.

IV. REFRAMING THE SCOPE OF HUMAN RESPONSIBILITY FOR AI ERRORS

The correlation between risk and responsibility has been deemed in the aforementioned argument. Those who enjoy the benefits should also bear the risks brought by their benefits. This argument opens up new questions of whether the scope

---

[116] Haleem, Abid, et al. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*. 2022, 3: 119-132.

[117] Wang, Yan. Do not go gentle into that good night: The European Union's and China's different approaches to the extraterritorial application of artificial intelligence laws and regulations. *Computer Law & Security Review*. 2024, 53: 105965.

of human responsibility varies according to the risk level of AI. This question has not yet been fully discussed in the legal scholarship. In order to fill the gap in the discussion of responsibility within this field, the chapter first outlines a wide array of obligations of human subjects responsible for AI, based on the risk levels classified in the AI Act. Second, I argue that the scope of human responsibility should vary according to the risk level of AI. Those responsible for high-risk AI products should bear additional obligations, namely the obligation to provide technical authentication.

A. Joint Obligations for Human Subjects Responsible for AI Across Different Risk Levels

The AI Act acknowledges that the degree of autonomy in AI products affects their risk levels and classifies them into four categories: unacceptable risk, high risk, limited risk, and minimal risk.[118] According to the AI Act, the table below illustrates the types of AI products at each risk level, along with the associated obligations of the entities responsible for them.

Level of Risk	Examples of AI Products	Obligations for Human Subjects Responsible for AI
Unacceptable risk	Real-time remote biometric recognition and social scoring in public spaces for law enforcement purposes	The use of these products is strictly restricted as they conflict with EU values. They could manipulate individuals and cause physical or psychological harm to the biometric identity system.[119] Such products may only be used under stringent conditions: for targeted searches for victims, preventing terrorist attacks or imminent threats to life, or tracking suspects or perpetrators of serious crimes.[120]
High risk	(a) Critical Infrastructure: Systems that could jeopardize citizens' lives and health. (b) Education and Vocational Training: Tools that impact educational opportunities and career prospects, such as automatically scored exams.	In line with the HLEG AI ethics guidelines, the White Paper specifies that high-risk AI applications should adhere to key requirements centred on transparency, fairness, safety, and security. These requirements include:

[118] Veale, Michael, and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*. 2021, 22(4): 97-112.
[119] Edwards, Lilian. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*. 2021, 1.
[120] Id.

	(c) Employment and Worker Management: Systems used for automated recruiting and resume triage. (d) Essential Services: Automated welfare systems and private sector credit scoring. (e) Law Enforcement: Systems that may infringe on fundamental rights, such as automated risk scoring for bail, deepfake detection, and pre-crime detection. (f) Immigration and Border Control: Tools for verifying travel documents and processing visas. (g) Judicial and Democratic Processes: Automated sentencing assistance and "robo-justice" systems.[121]	(a) Training Data: Ensuring data quality and relevance. (b) Data and Record-Keeping: Maintaining thorough documentation and data management practices. (c) Information to be Provided: Clearly communicating relevant information about the AI system. (d) Robustness and Accuracy: Ensuring the system performs reliably and accurately. (e) Human Oversight: Implementing mechanisms for human intervention and oversight. (f) Specific Requirements: Addressing unique considerations for certain applications, such as remote biometric identification.[122]
Limited risk	Products with limited risks include chatbots and emotion recognition systems.	The human subjects responsible for these products should fulfil obligations related to transparency, information disclosure, and explanation. For instance, providers of chatbots must clearly inform users that they are interacting with machines rather than humans.[123]
Minimal risk.	Simple chatbots or rule-based recommendation systems.	no special regulatory measures are required for AI products with minimal impact on users and society.[124]

This paper outlines the responsibilities of the individual with direct beneficiaries of the AI product as follows:

(1) Information Disclosure Obligation: Clearly indicating when information is generated by AI.

[121] Id.
[122] Id.
[123]Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.
[124] Id.

(2) Explanation Obligation: Providing detailed explanations about the AI's operations and outputs.

(3) Transparency Obligation: Disclosing the source and process of information generation.

(4) Monitoring and Reporting Obligation: Regularly monitoring the AI system and reporting relevant issues.

To elaborate, when people recognize that they are interacting with AI rather than humans, they perceive a difference in the credibility of AI responses compared to human responses. Human interaction often fosters a sense of intimacy, which enhances trust in the discourse.[125] The transparency and explanation obligations help beneficiaries verify whether the results of AI are correct and can be explained and therefore reasonable. Additionally, the information disclosure obligation helps set accurate expectations about the authenticity of AI products. This enables beneficiaries to detect anomalies promptly and trace the causes of errors.[126] For instance, if a product is disclosed as being generated by a deepfake system, beneficiaries will recognize it as fake and will not mistakenly believe it to be real. Moreover, the monitoring and reporting obligations require the responsible party to promptly report any abnormalities or potential risks to the relevant regulatory authorities and implement necessary corrective measures.[127]

B. Obligations of Human Subjects Responsible for High-Risk AI Products to Provide Technical Authentication

I argue that the persons in charge of high-risk AI products should bear obligations beyond the obligations outlined above. I focus my argument on the context of the deepfake evidence, as a typical example of high-risk AI products. The analysis is conducted within the judicial context to underscore the importance of the obligation to provide technical authentication. Deepfake evidence refers to false evidence generated by utilising deepfake technology. The term 'Deepfake defence' refers to the assertion by defence lawyers that the evidence in question has been fabricated using deepfake technology.[128] For instance, Reffitt, an alleged member of the anti-government group "Three Percenters," travelled from Texas to attend the pro-Trump rally in Washington, DC.[129] Video footage showed Reffitt in riot gear, carrying a gun, leading the crowd, and directing the attack on the Capitol. The defence team, led by Reffitt's legal counsel, presented their case to the jury, asserting that the video and image evidence were, in fact, deepfakes. The deepfake defence can be highly effective because the technical features of deepfakes, which rely on deep learning and generative adversarial networks, make it challenging to discern the authenticity of the evidence.

---

[125] Smuha, Nathalie A., et al. How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an Artificial Intelligence Act. *Available at SSRN 3899991*. 2021.
[126] Id.
[127] Oduro, Serena, Emanuel Moss, and Jacob Metcalf. Obligations to assess: Recent trends in AI accountability regulations. *Patterns*. 2022, 3(11).
[128] Dalila Durães, Pedro Miguel Freitas & Paulo Novais, *The Relevance of Deepfakes in the Administration of Criminal Justice* in *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (Springer International Publishing 2023) 351-369.
[129] Delfino, Rebecca A. The Deepfake Defense-Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers. *Ohio St. LJ*. 2023, 84: 1067.

Consequently, the deepfake defence is difficult to refute and can potentially favour the defendant.

Due to the advanced technology behind deepfakes, images and videos created by deepfake software can be indistinguishable from real ones, effectively creating convincing yet false content.[130] This realistic effect disrupts the cognitive logic that "seeing is believing and hearing is not." The deepfake defence not only challenges the fact-finder's cognitive belief that "seeing is believing," but may also lead to broader scepticism about the authenticity of all content.[131] This could result in doubts about unaltered evidence, as people might suspect it could be a realistic deepfake. A further significant concern is the possibility of a "liar's dividend" in the context of deepfake defence.[132] This term refers to a situation where the defence lawyer, aware that the evidence is genuine, still argues that it could be fake, exploiting the uncertainty surrounding deepfakes.[133] The "liar's dividend" encourages defence lawyers to employ deepfake defence, leading fact-finders to doubt or even disbelieve genuine evidence. This incentive further motivates defence lawyers to persist in using this strategy. For example, a deepfake defence might be used strategically to prevent the other party from participating in the lawsuit.[134] Conversely, proving or disproving deepfakes can require expensive expert fees, which some parties may not be able to afford. Some litigants may be unable to initiate or defend against lawsuits involving deepfake evidence due to the high cost of proving or disproving it. This could result in repeated success for the deepfake defence.

It has been observed in the Reffitt case, that Reffitt's lawyer presented only a suspicion without providing preliminary evidence to support the claim.[135] Such preliminary evidence includes, for example, information provided by the lawyer indicating that the source of the forged material is unknown or untrustworthy and cannot be traced back to a reliable distribution channel. To mitigate the "liar's dividend" associated with deepfake defence and to protect the principle of "seeing is believing," it has been suggested that defence lawyers should be restricted from arbitrarily raising deepfake claims. Conditions for presenting a deepfake defence should be strictly regulated. A more immediately efficacious approach might be to steer the technology from a technical standpoint. Judges should require defence lawyers to demonstrate a good-faith basis for alleging that evidence is a deepfake and conduct technical authentication of such evidence during pre-court meetings.[136] If it is challenging to determine whether evidence has been tampered with using deepfake technology, it is often deemed inadmissible. For instance, in *People v. Beckley*, the appellate court rejected the prosecution's request to admit a photo because neither experts nor fact

---

[130] Pfefferkorn, Riana. "Deepfakes" in the Courtroom. *BU Pub. Int. LJ*. 2019, 29: 245-276.
[131] LaMonaca, John P. "A break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes." *Am. UL Rev.* 69 (2019): 1945.
[132] Delfino, Rebecca. Pay-to-play: Access to Justice in the Era of AI and Deepfakes. *Available at SSRN 4722364*. 2024.
[133] Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natália S. Bueno. The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?. *American Political Science Review*. 2023: 1-20.
[134] Buckland, Robert. AI, Judges and Judgment: Setting the Scene. *M-RCBG Associate Working Paper Series*. 2023.
[135] Delfino, Rebecca A. Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review*. 2019, 88(3): 887-938.
[136] Palmiotto, Francesca. Detecting deep fake evidence with artificial intelligence: A critical look from a criminal law perspective. *Available at SSRN 4384122*. 2023.

witnesses could authenticate its authenticity.[137] Additionally, the website where the photo was posted did not monitor or independently verify its authenticity. Due to the ease with which the photo could be tampered with, the court ultimately deemed it inadmissible.[138]

Authentication is a method from the Anglo-American legal system used to assess the authenticity of evidence and establish the specific facts of a case.[139] Due to the virtuality, separability, and volume of electronic data, determining its authenticity is challenging when relying solely on visual inspection without technical identification methods.[140] Consequently, the rise of electronic data has led to the evolution of authentication methods from traditional approaches to technical methods, including data integrity verification, trusted timestamps, and digital signatures. Technical authentication methods have advanced significantly in recent years, offering robust support for identifying high-risk AI products.[141] Deepfake detection tools encompass both image and video detection models. Image detection models, for example, use deep convolutional neural networks to identify fake images generated by generative adversarial networks, employing techniques like Gaussian blur and noise to detect alterations in human pictures.[142] Video detection models include methods for capturing facial forgeries, analyzing timers of deepfakes, and examining audio-video relationships. These models detect fake videos by analyzing physical properties like pulsation and extracting features from frames using convolutional neural networks.[143] For instance, the app called *'eyeWitness to Atrocities' ca*n provide information about when and where a photo or video was taken. It helps to verify its authenticity and ensure it has not been tampered with. The app's transmission protocol and secure server system establish a chain of custody, thereby enabling the integrity of the information to be maintained.[144]

Since technical identification methods are typically controlled by AI developers, defence lawyers often face significant challenges in accessing these methods independently. This lack of access can hinder the defence's ability to effectively challenge the authenticity of evidence presented against their clients, particularly in cases involving deepfake technology. To mitigate this issue, it is essential for deepfake

---

[137] Delfino, Rebecca A. Deepfakes on trial: a call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *Hastings LJ*. 2022, 74: 293-348.
[138] Mehlman, Julia. Facebook and myspace in the courtroom: authentication of social networking websites. *Criminal Law Brief*. 2012, 8(1): 9-28.
[139] MacNeil, Heather, and Heather MacNeil. Trusting Records as Legal Evidence: Common Law Rules of Evidence. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. 2000: 32-56.
[140] Mnookin, Jennifer L. Scripting expertise: The history of handwriting identification evidence and the judicial construction of reliability. *Virginia Law Review*. 2001: 1723-1845.
[141] Liang, Weixin, et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*. 2022, 4(8): 669-677.
[142] Jones, Karl, and Bethan Jones. How robust is the United Kingdom justice system against the advance of deepfake audio and video?. *Electrotechnica and Electronica (E+E)*. 2022, 57 (9-12): 103-109.
[143] Durães, Dalila, Pedro Miguel Freitas, and Paulo Novais. The relevance of deepfakes in the administration of criminal justice. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. 2023, 351-369.
[144]Gregory, Sam. Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*. 2022, 23(3): 708-729.

technology developers to facilitate greater access for defence lawyers.[145] This would necessitate cooperation from AI developers, who should ensure AI transparency by providing identification tools and methodologies for independent scrutiny to support the legal process. Moreover, current deepfake detection technology often lags behind production technology, creating a significant gap that can be exploited in legal contexts.[146] By mandating that responsible entities provide reliable identification of deepfake materials, legal frameworks can incentivize AI development companies to enhance their detection capabilities. Such requirements would not only foster innovation in detection technology but also help ensure that the justice system can effectively address the challenges posed by AI errors.

CONCLUSION

This essay concludes by suggesting that humans cannot establish a trust relationship with AI. This implies that AI could not be expected to take responsibility. This essay aligns with the prevailing perspective within the legal scholarship, which holds that only humans should be responsible for AI errors. It seeks to rectify a misconception that the reliability of AI should not be conflated with its trustworthiness. The development of "trustworthy AI" necessitates that the human subject responsible for AI overcome the opacity and lack of explainability of AI in technology. In addition, there is a need for a shift in focus from reliability to the trustworthiness of AI. Nevertheless, the term "trustworthy AI" is not a tangible trust relationship between humans and AI, but rather a sense of trust generated by the appreciation of AI's behaviour. The lack of intrinsic moral motivation and accountability inherent to AI makes it challenging to cultivate genuine trust. The lack of moral responsibility inherent to AI precludes its potential to become a responsible subject. The term "trustworthy AI" should be understood to mean "a trustworthy human subject who is responsible for AI errors." This implies that we trust humans who possess the capacity to regulate AI. In light of the above, this essay reframes the scope of responsibility of human subjects. It is posited that only those who are direct beneficiaries of the AI product should bear responsibility. Furthermore, this essay stresses the importance of differentiating the obligations of these stakeholders in line with the risk level of AI, particularly those of high-risk AI applications. These applications must adhere to the AI Act and fulfil obligations for technical authentication.

This essay contributes to the field by applying the virtue jurisprudence-based approach to the issue of reliability and trustworthiness in AI. It highlights the distinction between these terms, indicating that reliability is not a prerequisite for trust. This approach demonstrates why AI products can engender a sense of trust in people, given that AI is capable of imitating human virtuous behaviour. Virtue jurisprudence posits that establishing a trust relationship requires one subject to know and believe that the moral behaviour of another is based on genuine moral motivations and that this party can be held accountable for any mistakes. Although AI can be programmed to exhibit virtuous behaviours and elicit emotional responses, it fundamentally lacks moral motivations and cannot grasp the moral significance of its actions. Additionally, AI lacks phronesis and cannot bear moral responsibility. Therefore, this approach explains

[145] Caldera, Elizabeth. Reject the evidence of your eyes and ears: deepfakes and the law of virtual replicants. *Seton Hall L. Rev.* 2019, 50: 177-206.
[146] LaMonaca, John P. A break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes. *Am. UL Rev.* 2019, 69: 1945-1988.

why humans cannot establish a trust relationship with AI. Moreover, this essay reframes the scope of subjects and objects of responsibility for AI errors. It combines human responsibility with direct benefits and risk levels, offering a comprehensive approach to this complex issue.

It is important to underline that for AI products with varying levels of risk, the distinctions in the scope of responsibility of the designated individual are primarily addressed from the perspective of judicial evidence and proof. This focus is concentrated on the challenges posed by high-risk AI products. While the ongoing perfusion of technology into the legal system may be inexorable, this essay aims to encourage further scholarly attention to the previously unexamined question of whether the scope of human responsibility varies according to the risk level of AI. Future research should explore the differences in obligations among individuals at different risk levels across various fields.

BIBLIOGRAPHY

Regulation:

1. UK, GOV. Ethics, transparency and accountability framework for automated decision-making. 2021.

Books:

2. Shannon Vallor. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) ch. 1.

3. Linda Trinkaus Zagzebski. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge* (Cambridge University Press 1996) ch. 1.

Articles:

1. Brayne, Sarah. The criminal law and law enforcement implications of big data. *Annual Review of Law and Social Science*. 2018, 14(1): 293-308.

2. Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.

3. Lei, Cheng. Legal control over Big Data criminal investigation. *Social Sciences in China*. 2019, 40(3): 189-204.

4. Crawford, Kate, and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.* 2014, 55: 93-128.

5. Grimm, Paul W., Maura R. Grossman, and Gordon V. Cormack. Artificial intelligence as evidence. *Nw. J. Tech. & Intell. Prop.* 2021, 19: 9-106.

6. Roth, Andrea. Machine Testimony. *Yale Law Journal*. 2017, 126: 1972.

7. Phelps, Kaelyn. Pleading Guilty to Innocence: How Faulty Field Tests Provide False Evidence of Guilt. *Roger Williams UL Rev.* 2019, 24: 143-166.

8. Lior, Anat. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 2019, 46: 1043-1102.

9. Jiang, Binxiang. Research on factor space engineering and application of evidence factor mining in evidence-based reconstruction. *Annals of Data Science*, 2022, 9(3): 503-537.

10. Katyal, Sonia K. Private accountability in the age of artificial intelligence. *UCLA L. Rev.* 2019, 66: 54-141.

11. Rodriguez, Xavier. Artificial Intelligence (AI) and the Practice of Law in Texas. *S. Tex. L. Rev.* 2023, 63: 1-35.

12. Manojkumar, P. K., et al. AI-based virtual assistant using python: a systematic review. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2023, 11: 814-818.

13. Rosenberg, Louis. The manipulation problem: conversational AI as a threat to epistemic agency. *arXiv preprint arXiv:2306.11748*. 2023.

14. Taye, Mohammad Mustafa. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*. 2023, 12(5): 91.

15. Kushwah, Preeti. Evaluating the Evidential Value of Evidence Generated by AI. *Issue 6 Indian JL & Legal Rsch.* 2022, 4: 1-11.

16. Leone, Massimo. From fingers to faces: Visual semiotics and digital forensics. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*. 2021, 34(2): 579-599.

17. Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.

18. Grossman, Maura R., et al. The GPTJudge: justice in a generative AI world. *Duke Law & Technology Review*. 2023, 23(1): 1-26.

19. Chan, Janet, and Lyria Bennett Moses. Is big data challenging criminology?. *Theoretical criminology*. 2016, 20(1): 21-39.

20. Wheeler, Billy. Giving Robots a Voice: Testimony, Intentionality, and the Law. *Androids, Cyborgs, and Robots in Contemporary Culture and Society*. IGI Global, 2018. 1-34.

21. Schmidt, Philipp, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 2020, 29(4): 260-278.

22. Quezada-Tavárez, Katherine, Plixavra Vogiatzoglou, and Sofie Royer. Legal challenges in bringing AI evidence to the criminal courtroom. *New Journal of European Criminal Law*. 2021, 12(4): 531-551.

23. Lv, Zhihan. Generative artificial intelligence in the metaverse era. *Cognitive Robotics.* 2023, 3: 208-217.

24. Gless, Sabine. AI in the Courtroom: a comparative analysis of machine evidence in criminal trials. *Geo. J. Int'l L.* 2019, 51: 195-254.

25. Padovan, Paulo Henrique, Clarice Marinho Martins, and Chris Reed. Black is the new orange: how to determine AI liability. *Artificial Intelligence and Law*. 2023, 31(1): 133-167.

26. Hew, Patrick Chisan. Artificial moral agents are infeasible with foreseeable technologies. *Ethics and information technology*. 2014, 16: 197-206.

27. Hakli, Raul, and Pekka Mäkelä. Moral responsibility of robots and hybrid agents. *The Monist*. 2019, 102(2): 259-275.

28. Sahoh, Bukhoree, and Anant Choksuriwong. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review.*Journal of Ambient Intelligence and Humanized Computing.* 2023, 14(6): 7827-7843.

29. Freiman, Ori. Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs. *Episteme*. 2023: 1-17.

30. Shneiderman, Ben. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*. 2010, 10(4): 1-31.

31. Bryson, Joanna J., and Andreas Theodorou. How society can maintain human-centric artificial intelligence. *Human-centered digitalization and services*. 2019: 305-323.

32. Ulgen, Ozlem. A human-centric and lifecycle approach to legal responsibility for AI. *Communications Law Journal: Journal of Computer, Media and Telecommunications Law*. 2021, 26(2): 1-15.

33. Ho, Calvin Wai-Loon, and Karel Caals. "How the EU AI Act Seeks to Establish an Epistemic Environment of Trust." *Asian Bioethics Review* (2024): 1-28.

34. Fukuda-Parr, Sakiko, and Elizabeth Gibbons. Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy*. 2021, 12: 32-44.

35. Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.

36. Ryan, Mark. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. 2020, 26(5): 2749-2767.

37. Opderbeck, David W. Artificial Intelligence, Rights and the Virtues. *Washburn LJ*. 2020, 60: 445-474.

38. van der Veer, Sabine N., et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*. 2021, 28(10): 2128-2138.

39. Schoenherr, Jordan Richard, and Robert Thomson. When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. *IEEE Transactions on Technology and Society*. 2024.

40. Davis, Joshua P. Law without mind: AI, ethics, and jurisprudence. *Cal. WL Rev.* 2018, 55: 165-220.

41. Fowers, Blaine J., Jason S. Carroll, Nathan D. Leonhardt, and Bradford Cokelet. The emerging science of virtue. *Perspectives on Psychological Science*. 2021, 16(1): 118-147.

42. Hagendorff, Thilo. A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*. 2023, 35(3): 55.

43. Buruk, Banu, Perihan Elif Ekmekci, and Berna Arda. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*. 2020, 23(3): 387-399.

44. Floridi, Luciano, and Jeff W. Sanders. Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*. 2001, 3: 55-66.

45. Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.

46. Konstantinos, Kouroupis, and Evie Lambrou. CHATGPT–ANOTHER STEP TOWARDS THE DIGITAL ERA OR A THREAT TO FUNDAMENTAL RIGHTS AND FREEDOMS?. *Pravo-teorija i praksa*. 2023, 40(3): 1-18.

47. Skjuve, Marita, et al. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies.* 2021, 149: 102601.

48. Lukka, Kari, and Petri Suomala. Relevant interventionist research: balancing three intellectual virtues. *The Societal Relevance of Management Accounting*. Routledge, 2017: 132-148.

49. Constantinescu, Mihaela, et al. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. 2021, 23: 803-814.

50. Contini, Francesco. Artificial intelligence and the transformation of humans, law and technology interactions in judicial proceedings. *Law, Tech. & Hum.* 2020, 2: 4-18.

51. Amaya, Amalia. Virtuous adjudication; or the relevance of judicial character to legal interpretation. *Statute Law Review*. 2019, 40(1): 87-95.

52. Widłak, Tomasz. Judges' virtues and vices: outline of a research agenda for legal theory. *Archiwum Filozofii Prawa i Filozofii Społecznej*. 2019, 20(2): 51-62.

53. Brady, Michael S., and Duncan Pritchard. Moral and epistemic virtues. *Metaphilosophy*. 2003, 34(1/2): 1-11.

54. Stover, James, and Ronald Polansky. Moral virtue and megalopsychia. *Ancient Philosophy.* 2003, 23(2): 351-359.

55. Kristjánsson, Kristján, et al. Phronesis (practical wisdom) as a type of contextual integrative thinking. *Review of General Psychology*. 2021, 25(3): 239-257.

56. Tamò-Larrieux, Aurelia, et al. Regulating for trust: Can law establish trust in artificial intelligence?. *Regulation & Governance*. 2023.

57. Farina, Mirko, Petr Zhdanov, Artur Karimov, and Andrea Lavazza. AI and society: a virtue ethics approach. *AI & SOCIETY*. 2022: 1-14.

58. Sutrop, Margit. Should we trust artificial intelligence?. *Trames*. 2019, 23(4): 499-522.

59. Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.

60. Hallamaa, Jaana and Taina Kalliokoski. How AI systems challenge the conditions of moral agency?. *International Conference on Human-Computer Interaction*. 2020: 54-64.

61. Lentzas, Athanasios, and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review*. 2020, 53(3): 1975-2021.

62. Zhou, Yuanyuan, et al. How human–chatbot interaction impairs charitable giving: the role of moral judgment. *Journal of Business Ethics*. 2022, 178(3): 849-865.

63. Adamopoulou, Eleni, and Lefteris Moussiades. An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*. Springer, Cham, 2020: 373-383.

64. Wilson, Abigail, Courtney Stefanik, and Daniel B. Shank. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations?. *Computers in Human Behavior Reports*. 2022: 8: 100229.

65. Cave, S., Nyrup, R., Vold, K., & Weller, A. Motivations and risks of machine ethics. *Proceedings of the IEEE*. 2018, 107(3): 562-574.

66. Haladjian, H. H., & Montemayor, C. Artificial consciousness and the consciousness-attention dissociation. *Consciousness and Cognition*. 2016, 45: 210-225.

67. Grace, Katja, et al. Thousands of AI authors on the future of AI. *arXiv preprint arXiv: 2401. 02843*. 2024.

68. Solaiman, Sheikh M. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial intelligence and law*. 2017, 25: 155-179.

69. Dahiyat, Emad Abdel Rahim. Law and software agents: Are they "Agents" by the way?. *Artificial Intelligence and Law*. 2021, 29(1): 59-86.

70. Baum, Kevin, et al. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*. 2022, 35(1): 12.

71. Judd, David. Disentangling DeVries: A Manufacturer's Duty to Warn Against the Dangers of Third-Party Products. *La. L. Rev.* 2020, 81(1): 217-270.

72. Laux, Johann, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*. 2024, 18(1): 3-32.

73. Edwards, Lilian. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*. 2021, 1.

74. Madiega, Tambiama. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*. 2021.

75. Hacker, Philipp. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, innovation and technology*. 2021, 13(2): 257-301.

76. Marano, Pierpaolo, and Shu Li. Regulating robo-advisors in insurance distribution: Lessons from the insurance distribution directive and the ai act. *Risks*. 2023, 11(1): 12.

77. Hacker, Philipp. The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*. 2023, 51: 105871.

78. Hudig, Dirk. The Problem of Low and Uncertain Risks: Balancing Risks and Benefits. *European Journal of risk regulation*. 2012, 3(2): 157-160.

79. Haleem, Abid, et al. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*. 2022, 3: 119-132.

80. Wang, Yan. Do not go gentle into that good night: The European Union's and China's different approaches to the extraterritorial application of artificial intelligence laws and regulations. *Computer Law & Security Review*. 2024, 53: 105965.

81. Veale, Michael, and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*. 2021, 22(4): 97-112.

82. Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.

83. Smuha, Nathalie A., et al. How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an Artificial Intelligence Act. *Available at SSRN 3899991*. 2021.

84. Oduro, Serena, Emanuel Moss, and Jacob Metcalf. Obligations to assess: Recent trends in AI accountability regulations. *Patterns*. 2022, 3(11).

85. Dalila Durães, Pedro Miguel Freitas & Paulo Novais, *The Relevance of Deepfakes in the Administration of Criminal Justice* in *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (Springer International Publishing (2023) 351-369.

86. Delfino, Rebecca A. The Deepfake Defense-Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers. *Ohio St. LJ*. 2023, 84: 1067.

87. Pfefferkorn, Riana. " Deepfakes" in the Courtroom. *BU Pub. Int. LJ*. 2019, 29: 245-276.

88. LaMonaca, John P. "A break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes." *Am. UL Rev.* 69 (2019): 1945.

89. Delfino, Rebecca. Pay-to-play: Access to Justice in the Era of AI and Deepfakes. *Available at SSRN 4722364*. 2024.

90. Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natália S. Bueno. The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?. *American Political Science Review*. 2023: 1-20.

91. Buckland, Robert. AI, Judges and Judgment: Setting the Scene. *M-RCBG Associate Working Paper Series*. 2023.

92. Delfino, Rebecca A. Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review*. 2019, 88(3) :887-938.

93. Palmiotto, Francesca. Detecting deep fake evidence with artificial intelligence: A critical look from a criminal law perspective. *Available at SSRN 4384122*. 2023.

94. Delfino, Rebecca A. Deepfakes on trial: a call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *Hastings LJ*. 2022, 74: 293-348.

95. Mehlman, Julia. Facebook and myspace in the courtroom: authentication of social networking websites. *Criminal Law Brief*. 2012, 8(1): 9-28.

96. MacNeil, Heather, and Heather MacNeil. Trusting Records as Legal Evidence: Common Law Rules of Evidence. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. 2000: 32-56.

97. Mnookin, Jennifer L. Scripting expertise: The history of handwriting identification evidence and the judicial construction of reliability. *Virginia Law Review*. 2001: 1723-1845.

98. Liang, Weixin, et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*. 2022, 4(8): 669-677.

99. Jones, Karl, and Bethan Jones. How robust is the United Kingdom justice system against the advance of deepfake audio and video?. *Electrotechnica and Electronica (E+E)*. 2022, 57(9-12): 103-109.

100. Durães, Dalila, Pedro Miguel Freitas, and Paulo Novais. The relevance of deepfakes in the administration of criminal justice. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. 2023, 351-369.

101. Gregory, Sam. Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*. 2022, 23(3): 708-729.

102. Caldera, Elizabeth. Reject the evidence of your eyes and ears: deepfakes and the law of virtual replicants. *Seton Hall L. Rev.* 2019, 50: 177-206.

REIMAGINING LEGAL ETHICS: CO-EXISTENCE OF DOMINANT AND ALTERNATIVE PRINCIPLES IN LAWYERING

Anthonie Fan*

Abstract: The dominant model of legal professionalism requires lawyers to advocate zealously for their clients, regardless of personal moral objections. This role, stemming from the adversarial system, positions lawyers as resolute representatives who must prioritize client interests above all else. However, as instances where lawyers have justified or facilitated egregious acts come to light, there is growing resistance from the public and newer legal practitioners against the notion that unwavering client advocacy is a professional necessity. Critics like John Kennedy, Trevor Farrow, Robert Vischer, and Allan Hutchinson argue for a reevaluation of this antiquated ethical stance, advocating for a model that considers the broader societal impact of legal practice. This paper explores the ethical dilemmas of the dominant model, proposing a new regulatory framework that balances the merits of both dominant and alternative views. The proposed system divides the legal profession into three segments: Advocates, Juridical Scriveners, and In-House Advisors. Advocates, akin to traditional trial lawyers, would operate under a modified dominant model, employed and regulated by an independent body to ensure fair distribution of cases and mitigate financial pressures. Juridical Scriveners, responsible for transactional and advisory work, would have ethical guidelines encouraging moral reflection and accountability, acknowledging their role in shaping legal documents with significant social impact. In-House Advisors, balancing roles of both Advocates and Scriveners, would require tailored regulations to address their unique client-employer relationship. This trichotomy aims to create a sustainable legal profession that respects individual practitioner needs while maintaining ethical integrity and access to justice.

Keywords: Legal Ethics; Dominant Model; Sustainable Lawyering

* Trinity College, University of Toronto, Canada.

Table of Contents

Introduction 220

I. The Case for the Dominant Model 220

A. The Adversarial System 220

B. Access to Justice 221

II. Challenges to the Dominant Model 222

A. Lack of Consideration for Non-Partisan Interests 222

B. Appropriation and Misapplication 223

III. Proposed Restructuring of the Profession 224

A. Advocates 224

1. General Parameters 224

2. Compensation Scheme 224

3. Disciplinary Regime 225

4. Adherence to Individual Development 225

B. Juridical Scriveners 226

1. Case Study: England and Québec 226

a. English Barristers and Solicitors 226

b. Québec *Avocats* and *Notaires* 227

c. Comparison: *Notaires* vs. Solicitors 227

d. Comparison: *Avocats* vs. Barristers 228

e. Adversarial vs. Inquisitorial Ethics 229

2. Professional Parameters and Ethics of Juridical Scriveners 229

a. Areas of Practice 229

b. Professional Ethics 230

i. Moral Investment in Client's Cause 230
ii. Interests of Third Parties and Society 231
iii. Scrivener-Client Candor 232
iv. Conclusion 232

IV. Future Research 233

A. In-House Advisors 233

B. Compensation Scheme for Advocates 233

Conclusion 233

Bibliography 235

INTRODUCTION

Traditionally, lawyers have been conceived as "amoral technicians" whose loyalty lies only with their clients.[1] Under this view, otherwise known as the "dominant model,"[2] lawyers act as a resolute mouthpiece for their clients regardless of their personal views.[3] However, as cases where lawyers have justified or aided heinous acts committed by their clients are exposed,[4] the public and younger generations of legal practitioners alike have grown averse to the idea that a "pact with the Devil" is a necessary rite of passage for a career in law.[5] To address the ethical dilemmas posed by the dominant model, legal theorists like John Kennedy, Trevor Farrow, Robert Vischer, and Allan Hutchinson have called on the profession to exercise introspection and review the antiquated mantras of modern legal ethics.

This paper aims to enhance the sustainability of the legal profession by harmonizing the dominant model with aspects of its proposed alternatives. The paper will start by explaining the appeals of the dominant view by identifying the principles embedded in the dominant view. The paper will then discuss rightful objections to the dominant view raised by its critics before finding common ground between the dominant and alternative views of the legal profession. Finally, this paper will propose a new regulatory regime for the legal profession that embodies the merits of the dominant model and its proposed alternatives.

I. THE CASE FOR THE DOMINANT MODEL

The dominant view of professionalism requires a lawyer to advocate wholeheartedly for their client's position, regardless of how "distasteful" the position may be.[6] The role of lawyers as but a learned mouthpiece, coupled with solicitor-client confidentiality,[7] paints the relationship between a lawyer and their client as one secluded from the rest of the world. Indeed, Lord Henry Brougham famously proposed a romanticized mantra of the dominant model: A lawyer "knows but one person in all the world, and that person is [their] client."[8] Under this view, a good lawyer is one who places their clients above all else, including their own interests and the interests of any moral or political causes.[9]

A. The Adversarial System

The dominant model finds its philosophical backdrop in the adversarial legal system. In the adversarial system, parties are in charge of conducting their own litigation with minimal intervention from the judge.[10] Each party is expected to present

---

[1] Robert K. Vischer, "Legal Advice as Moral Perspective," SSRN Scholarly Paper (Rochester, NY, August 10, 2005), https://papers.ssrn.com/abstract=771006, 36.
[2] Vischer, 36.
[3] Vischer, 36; Trevor C. W. Farrow, "Sustainable Professionalism," *Osgoode Hall Law Journal* 46, no. 1 (January 1, 2008): 51–103, https://doi.org/10.60082/2817-5069.1207, 56-7.
[4] Vischer, "Legal Advice as Moral Perspective," 1-2.
[5] Farrow, "Sustainable Professionalism," 53-4.
[6] Farrow, 63-7.
[7] Farrow, 56n28.
[8] Lord Henry Brougham qtd. in Farrow, 63-4.
[9] Farrow, 57.
[10] Neil Brooks, "The Judge and the Adversary System," in *The Civil Litigation Process Cases and Materials*, ed. Janet Walker et al., 7th ed. (Toronto: Emond Montgomery Publications, 2010), 91.

all relevant factual and legal arguments before the court, whose power is limited to adjudicating disputes brought by the parties.[11] Under this system, the parties—rather than the court—are at the heart of the justice system.[12] This emphasis on party autonomy reinforces the sacrosanct relationship between a lawyer and their client.

One main argument used to support the adversarial system of justice is that it prioritizes the agency of parties.[13] This principle is reflected in the dominant model of legal ethics by allowing a litigant ultimate control over what is presented in court. Although the adversarial system, in its ideal form, welcomes all to advocate for their causes, the complexity of the legal system often makes effective advocacy a cumbersome task for an average citizen. A lawyer's role in the system, therefore, appears to be a professionally trained counselor who helps their clients find ways to exercise their autonomy as an individual and as a party to the litigation.[14] As such, advocacy for "distasteful" ideas is a necessary price to pay for living under a system where individuals retain full autonomy over the arrangement of their lives.

Another bedrock of the justification for the adversarial system lies in its venue: an impartial tribunal.[15] Because the judge in the adversarial system is a qualified individual who maintains a professional distance from both parties, the system is expected to produce correct judgments through the competition of rival ideas proposed by the parties.[16] This aligns with one of the foundational assumptions of modern economics: The optimal allocation of goods arises when all parties strive to advance their self-interest to the fullest.[17] Because an independent arbiter watches over the marketplace of arguments inside the courtroom, the negative effects of "distasteful" positions are—from a theoretical point of view—confined to the four walls of the courtroom.

The dominant model of professionalism, where lawyers are free to advocate any position they receive instructions on, arises from these two tenets of the adversarial system. Conversely, the adversarial legal tradition makes it difficult to divorce our judicial system entirely from the dominant model.

B. Access to Justice

In addition to the underlying design of the adversarial legal system, concerns with access to justice also warrant some support for the dominant model. Although an unlikely scenario, it raises bona fide concerns vis-à-vis access to justice if the last lawyer in town refuses to take a client on the basis of differences in opinion.[18] This issue is particularly manifest in criminal cases. Since the accused is facing prosecution by the State—who possesses more investigative, financial, and legal resources than an average citizen—and may lose their liberty as a result of the trial, it is a generally

---

[11] Brooks, 93-4.
[12] Brooks, 94.
[13] Brooks, 98-101.
[14] Farrow, "Sustainable Professionalism," 64-5.
[15] Brooks, "The Judge and the Adversary System," 103.
[16] Brooks, 103.
[17] Brooks, 99.
[18] Farrow, "Sustainable Professionalism," 92.

accepted maxim that they deserve the "best defense and representation possible," regardless of their character and the nature of their charges.[19]

I argue that the same principles that protect the dominant model in criminal cases extend to civil litigation as well. Although the State does not intervene in private matters between two subjects with no invitation,[20] the judgment made by the State's judicial organ is given the weight of law. Since the State can be engaged in the enforcement of adjudicated private matters, the lack of proper representation in private law litigation puts individuals in a vulnerable position when later faced with the aftermath of the trial. It would be a matter of fiction to say that the foreclosure of someone's only home and the repossession of their only car are categorically less important to their lives than spending a fortnight in jail. The dominant model, thus, has a legitimate place in private law litigation on grounds of access to justice as well.

II. CHALLENGES TO THE DOMINANT MODEL

A. Lack of Consideration for Non-Partisan Interests

One of the major setbacks of the dominant model shares its roots with the best arguments in favor of it—the adversarial legal system. In a classic courtroom setting, only two parties are present: the prosecution and the defense. The real-life problems underlying the legal disputes, however, rarely deal only with the interests of two parties.[21] The framework where the dominant model finds home is unfit to accommodate the interests of the lawyers, of the legal profession, of other affected parties, and of society at large.[22] Although third parties and public interest groups can sometimes find their way to a day in court, the adversarial system reserves the center stage exclusively for the two parties to litigation. This problem, however, is rooted in the misalignment of interests for the lawyers and for the courts.

Our current legal scene features lawyers practicing as competitors offering legal services to the public. Lawyers are unable to sustain themselves financially if they fail to maintain a reputation of being loyal to their clients and likely to win cases. The current professional landscape, therefore, incentivizes lawyers to only look at client interests, often at great expense to the image of the legal profession and to the lawyer's own developmental needs.[23] For similar reasons, many litigators dispense patently overboard litigation strategies with the sole purpose of burdening the other party and driving them out of participation.[24] These practices come at the expense of the public interest as well. Nevertheless, the reputation of the profession, the needs of the lawyer as a socialized person, and the needs of society are hardly relevant when caring for these needs might cost one their source of income.

Judges and jurors, on the other hand, are being forced into having tunnel vision by the mainstream narrative of litigation that the only interests relevant to adjudication

---

[19] Farrow, 65.
[20] Brooks, "The Judge and the Adversary System," 95.
[21] Farrow, "Sustainable Professionalism," 87.
[22] Farrow, 87.
[23] Farrow, 89-90.
[24] Allan C. Hutchinson, "Fighting Fair - A Call to Ethical Arms," *The Professional Lawyer* 23, no. 3 (2016): 12–16.

are those presented in court.[25] It is particularly hard to escape this tunnel vision created by societal narrative when lawyers from both sides try to limit the discourse only to the facts favorable to their clients.[26]

These problems, however, do not mitigate the centrality of the dominant model to the adversarial system of justice. It remains highly unfeasible to dismantle the dominant model in the courtroom without destroying the very courtroom setting itself. Therefore, solutions to the neglect of diverse interests should come from a reform of the incentive structure of litigators.

B. Appropriation and Misapplication

Although the dominant model is closely intertwined with the adversarial system, not all lawyers specialize in courtroom advocacy. The safeguards against "distasteful" positions held by clients spilling over into society are not present in transactional and advisory matters. Because there is no opposing counsel pushing back on articles of incorporation nor an independent arbiter throwing out egregiously unfair distribution of an estate, the lawyers become the gatekeepers of justice in these situations. Similarly, when a lawyer drafts a memorandum justifying cruel and unusual punishments for the State,[27] they are not countered by judges or with civil rights lawyers that would otherwise serve as checks and balances on the legal opinion.

In these situations, a lawyer is more than a legal technician.[28] In private law matters, the shaper of deals is also a legislator[29] in the sense that the very document they prepare will serve as the basis for future litigation. Although proponents of the dominant model eschew this conception of lawyers,[30] law practitioners are the judge, the jury, the opposing counsel, and the legislature in transactional and advisory matters. As Kennedy points out, there is no room for the façade of neutrality when the lawyer is molding rules with their bare hands.[31]

The dominant model, however, provides a scapegoat to the lawyers willingly drafting morally reprehensible legal opinions.[32] This blame-shifting initiates a vicious cycle where the name of the profession becomes ever more tarnished as transactional and advisory lawyers appropriate and abuse the moral protection extended to their litigator counterparts. The less reputable the profession, however, the more justified it is for its members to misbehave and blame it on the industry.

The professional regulators, however, do not seem to have come to terms with this factual division in roles that lawyers play. Canadian and American jurisdictions feature a uniform set of ethical rules for all practitioners.[33] The evils of protecting "distasteful" legal craft in transactional and advisory operations, however, are not

---

[25] Brooks, "The Judge and the Adversary System," 98-100.
[26] Farrow, "Sustainable Professionalism," 84.
[27] Vischer, "Legal Advice as Moral Perspective," 1-2.
[28] Duncan Kennedy, "The Responsibility of Lawyers for the Justice of Their Causes," 1987, https://hdl.handle.net/2346/87006, 1160.
[29] Kennedy, 1160.
[30] Farrow, "Sustainable Professionalism," 64-5.
[31] Kennedy, "The Responsibility of Lawyers for the Justice of Their Causes," 1160.
[32] Vischer, "Legal Advice as Moral Perspective," 1-3.
[33] Farrow, "Sustainable Professionalism," 66-7.

outweighed by the philosophical underpinnings of the adversarial system. I argue that the real culprit is the appropriation and misapplication of the dominant model rather than the model itself.

III. PROPOSED RESTRUCTURING OF THE PROFESSION

Because the nature of legal work is diverse in the real world, a sustainable regulatory scheme should account for—instead of willfully ignoring—this reality. I propose a legal profession divided into three segments: Advocates, Juridical Scriveners, and In-House Advisors. I have chosen these terms with the intention of not borrowing any existing concept in the Canadian legal tradition, lest there be any confusion as to the content of each segment's work. Different rules will apply to each segment of the profession that corresponds with the nature of their work and the status they play both in the legal and societal landscapes. In this paper, I will focus on the role and regulatory scheme for Advocates, which will be based on a modified version of the dominant model.

A. Advocates

1. General Parameters

Under my proposed system, Advocates are lawyers specializing in conducting trials and appellate hearings. They appear exclusively on behalf of a party to the legal matter before an adjudicator. Although no type of adjudication is excluded from this definition, *only* Advocates are allowed to speak on behalf of a party. Under this system, an Advocate becomes a resolute warrior of their client in the purest form. Although this segment of the profession bears some semblance to English barristers, Advocates can take instructions either directly from clients or from other lawyers. This design is to ensure that the legal system does not become more cumbersome for people seeking remedies. Advocates will remain largely under the command of the dominant model of professionalism since their work is conducted entirely in the adversarial arena.

2. Compensation Scheme

Although the dominant model applies philosophically to the adversarial courtroom, pragmatic problems of unethical competition exist. One of the ways that the proposed system seeks to optimize the dominant model is by revamping the compensation scheme for Advocates. As discussed above, the current system of litigators being competitors in the same market encourages vicious competition at the expense of professional collegiality and the public interest. Under my proposed system, all Advocates will be salaried employees of an independent body that assigns cases, at random, to Advocates with relevant experience. Because Advocates are salaried, they are no longer under the pressure of billable hours or the amount of dispute.

Furthermore, because clients are assigned at random, Advocates are not subject to financial pressure from their clients (in the United States and in Canada) or from solicitors (in England). Since an Advocate's career advancement is no longer tied to their comparative reputation, there is no incentive to use shady means to "get ahead." Therefore, Advocates are able to perform their duties as true technicians of advocacy.

Finally, the employer of Advocates will charge clients rates that are geared to their income level and the amount of dispute. This will make sure that justice remains accessible to the people and help the professional body of Advocates be financially sustainable.

3. Disciplinary Regime

Another main cause of unethical competition is the low price of non-compliance. As illustrated in cases like *Groia v. Law Society of Upper Canada*,[34] professional regulators and courts are reluctant to step in and correct litigators from dubious practices. This low cost of infractions often leads to a false sentiment that "all is fair in love and war." As Hutchinson observes, the fact that a lawyer *is able to* effect a certain maneuver does not imply that they *ought to* do it.[35] On the other hand, because litigators do not share an employer, regulators are limited to suspending, limiting, or revoking a lawyer's ability to practice.[36] The severe nature of the punishments does not seem applicable to minor infractions of litigation ethics, leading to infractions being normalized.

Under the proposed system, however, the employer of Advocates can assume some level of disciplinary control in the form of pay deductions and other intermediate punishments available to corporate employers. Because the employer is the sole provider of Advocate services, it is driven by self-interest to maintain a good reputation for the profession, which leads to strong incentives to pursue internal disciplinary actions. This new internal disciplinary system should feature elements of military ethics as outlined by Hutchinson, such as proportionality and considerations of necessity. The disciplinary scheme, thus, incentivizes Advocates to internalize professional and public interests in their practice.[37]

4. Adherence to Individual Development

Finally, the employer under the proposed system will respect the individual will of practitioners. There are two layers to this claim. The first is that Advocates are free to choose their specialization while working for the general employer. Because the assignment of cases is based on a random draw among lawyers who have experience in the relevant legal field, it is up to individual practitioners to develop their niche. The personal circumstances and aspirations of lawyers cannot be overlooked, as a profession that systemically dissociates its members from their senses of self is not sustainable nor desirable.[38] Because Advocates are valued as individuals, they are more empowered to engage in critical reflections on their litigation strategies and their development as a lawyer.

Moreover, lawyers freely choose to become an Advocate under the proposed system. There are always going to be aspiring lawyers who resolutely embrace the lures of the adversarial system and the dominant model that comes with it.[39] Becoming an Advocate is a sustainable route of professional development for them. For those whose

---

[34] Groia v. Law Society of Upper Canada, 2018 SCC 27, accessed December 1, 2023.
[35] Hutchinson, "Fighting Fair - A Call to Ethical Arms."
[36] Ontario, "Law Society Act," R.S.O. 1990 c. L.8, s. 35.
[37] Farrow, "Sustainable Professionalism," 89-96.
[38] Farrow, 99.
[39] Farrow, 100.

personal ethical goals gravitate towards only taking specific clients or only making money, however, the proposed system gives them room to thrive in the other two segments of the profession. The professions of Juridical Scrivener and In-House Advisor would both allow more liberty in terms of the financial and ethical considerations involved in the practice of law. As Farrow comments, a sustainable legal landscape is not one where all values of the dominant model are subverted but one where different developmental needs of legal practitioners can be accommodated.[40]

B. Juridical Scriveners

1. Case Study: England and Québec

a. English Barristers and Solicitors

The United Kingdom has three independent legal systems: England and Wales, Scotland, and Northern Ireland. Each legal system retains a distinct set of professional regulators that oversee the legal profession. This section will look at the legal profession as it is structured and regulated in England and Wales. Unlike in most Canadian jurisdictions, barristers and solicitors are two separate and independent professions in England and Wales. An English lawyer must choose between practicing exclusively as a barrister or a solicitor. As a result, there is a clear professional boundary between the work done by barristers and that done by solicitors.

Traditionally, English solicitors draft instruments of conveyance and provide general legal advice to their clients. While conducting their main duties as transactional lawyers, solicitors in England may appear before subordinate courts to assist their clients in making simple appearances before a justice of the peace. Nevertheless, English solicitors are generally barred from conducting litigation since they do not hold the right of audience before senior courts. If litigants need trial representation, they must hire a barrister.[41] Generally speaking, members of the public can *only* hire a barrister *through* their solicitor, making it imperative to instruct two lawyers at the same time.[42]

Although the barrister-solicitor distinction appears to resolve the tension between transactional lawyers operating under rules meant for trial lawyers, the reality in England provides no satisfactory answer to sustainable professionalism in the United States and Canada. There are two major flaws in the English system *vis-à-vis* sustainable lawyering: the hegemony of adversarial ethics in regulating solicitors and the obstruction of access to justice.

Although England's legal profession recognizes the distinction between trial lawyers and transactional lawyers by requiring practitioners to choose between one and the other, regulators of both professions operate under the set of ethical principles derived from the trial process. Indeed, solicitors are required to act "in the best interests of each client."[43] The wording used by the Solicitors Regulation Authority mirrors

---

[40] Farrow, 98-100.
[41] Pye Tait Consulting and Bar Standards Board, "Provision of Legal Services by Barristers," May 2017, https://www.barstandardsboard.org.uk/asset/C551BDF1-4C2E-404C-B7A11B802F31A2C3/, 8.
[42] Pye Tait Consulting and Bar Standards Board, 8.
[43] Solicitors Regulation Authority, "SRA Standards and Regulations" (2019), https://www.sra.org.uk/solicitors/standards-regulations/, Principle 7.

precisely the language used to regulate barristers.[44] In other words, although English solicitors undertake transactional work exclusively, they are nevertheless accorded the same professional shield from moral scrutiny. The English system, therefore, leaves the door open for transactional lawyers to co-opt and appropriate adversarial ideals to justify morally reprehensible actions.

Another important lesson from studying the English system is that a restructured legal profession must not be set up in a way that deprives trial lawyers of access to their clients. A direct consequence of such a system is that instructing two lawyers at the same time causes financial and emotional strain on litigants.[45] There is, however, also an alarming trend that barristers tend to keep their instructing solicitors more informed than their clients.[46] Both issues create substantive barriers to justice since the end consumer of legal representation is being pushed out of the core decision-making circle.[47]

b. Québec *Avocats* and *Notaires*

Unlike other provinces and territories, Québec has two law societies: the *Barreau du Québec*, which regulates *avocats*, and the *Chambre des notaires du Québec*, which regulates *notaires*.[48] There are some ostensible similarities between the English and Québec systems at first glance. Like English barristers, Québec *avocats* have the exclusive right to appear before a court on behalf of a client.[49] Furthermore, only *avocats*—and not *notaires*—are allowed to conduct litigation by filing documents related to court proceedings.[50] Similar to English solicitors, Québec *notaires* can prepare conveyances,[51] draft articles of incorporation,[52] issue letters,[53] give legal advice,[54] and administer affidavits.[55] While English solicitors retain a limited right of appearance before subordinate courts for simple matters, Québec *notaires* can only represent their clients in non-contentious private law proceedings.[56] Although the *avocat-notaire* distinction bears semblance with the solicitor-barrister divide, Québec's regulatory scheme provides for two distinct sets of professional ethics for legal practitioners.

c. Comparison: *Notaires* vs. Solicitors

---

[44] Bar Standards Board, "The BSB Handbook" (2023), https://www.barstandardsboard.org.uk/for-barristers/bsb-handbook-and-code-guidance/the-bsb-handbook.html, Core Duty 2.
[45] Pye Tait Consulting and Bar Standards Board, "Provision of Legal Services by Barristers," 20.
[46] *Ibid* at 20.
[47] Although the Bar Council recently started a pilot program allowing barristers practicing in certain areas the ability to be retained by clients directly, most cases are still handled under the barrister-solicitor model.
[48] Although Québec law provides an official English translation for the terms *avocat* and *notaire*, the words "notary" and "advocate" are avoided because notary has a common-law meaning that is distinct, and advocate is used as a part of the system proposed in this paper.
[49] Québec, "Act Respecting the Barreau du Québec," CQLR c B-3, s. 128(2)(*a*).
[50] Québec, s. 128(1)(*b*).
[51] Québec, "Notaries Act," CQLR c N-3, s. 15(2).
[52] Québec, s. 15(3).
[53] Québec, s. 15(6).
[54] Québec, s. 15(5).
[55] Québec, ss. 15(1) and 9.
[56] Québec, s. 15(7).

In Québec, a *notaire* is a "public officer" who must consider the interests of *all* parties, including those of society at large.[57] The primary missions of a Québec *notaire* are defined as to "serve the public interest" and look out for "the good repute of the profession."[58] In fact, Québec's *Notaries Act* provides that a *notaire* is vested with "public authority" and must "act impartially" towards their clients.[59] In practical terms, this means that a Québec *notaire* must consider the interests of all parties that would be impacted when drafting legal instruments, even if the concerned parties may not appear before the *notaire*.[60] When offering legal advice, Québec *notaires* are required to act as "disinterested, frank, and honest" legal advisors rather than partisan counselors for their clients.[61]

As such, if a person wishes to set up a trust that would funnel all of his wealth into his lover's control and disallow his partner from accessing the money, a Québec *notaire* would have the professional obligation to advise the client against the decision. If the client should insist, the *notaire*'s duty to act with "probity, objectivity, and integrity" would likely preclude them from drafting the requested instruments.[62] Similarly, if a company hires a *notaire* for assistance in evading taxes, the *notaire* must consider the negative image that their complicity would cause to the ranks of *notaires* and not just their personal financial interests. Moreover, the argument that "if I don't do it, someone else will"[63] is rendered futile under the professional ethics governing *notaires*—since everyone is responsible for the entire profession's reputation, no one would be that *someone else* that follows through with morally questionable acts.

d. Comparison: *Avocats* vs. Barristers

The role of a Québec *avocat* is not dissimilar to that of a trial lawyer under the dominant model: Québec law provides that *avocats* are "officer[s] of the court"[64] and that they "must act at all times in the best interests of the client."[65] As legal professionals, Québec *avocats* have the duty to discourage clients from engaging in frivolous legal strategies that would result in an abuse of judicial resources. Should a client insist, the *avocat* must prioritize their allegiance to the court before that to their client and refuse to comply.[66] This duty applies equally to counsel in common law jurisdictions, where frivolous and vexatious claims invite sanctions. Otherwise than this formalistic constraint, however, Québec *avocats* are free to "act for a client no matter what [the *avocat*'s] opinion may be on the client's guilt or liability."[67] Notably, Québec *avocats* are allowed wide discretion to refuse acting for a particular client.[68] This freedom is not enjoyed by English barristers, who must accept instructions barring few exceptions.[69] However, as long as a Québec *avocat* retains their mandate from a client,

---

[57] Éducaloi, "Notaire," Éducaloi, accessed January 27, 2024, https://educaloi.qc.ca/capsules/notaire/.
[58] Québec, "Code of Ethics of Notaries," CQLR c N-3, r 2, s. 1.
[59] Québec, Notaries Act, s. 11.
[60] Éducaloi, "Notaire."
[61] Québec, Code of ethics of notaries, s. 7.
[62] Québec, s. 13.
[63] Kennedy, "The Responsibility of Lawyers for the Justice of Their Causes," 1160.
[64] Québec, Act respecting the Barreau du Québec, s. 2.
[65] Québec, "Code of Professional Conduct of Lawyers," CQLR c B-1, r 3.1, s. 23.
[66] Québec, s. 41.
[67] Québec, s. 32.
[68] Québec, s. 33.
[69] Bar Standards Board, The BSB Handbook, rC29.

it must act resolutely in the interest of that client, no matter their personal opinion about the client's cause.[70]

Although Québec *avocats* who practice as trial lawyers have similar professional obligations and expectations as their English counterparts, current Québec law also allows *avocats* to practice in transactional matters. In fact, Québec's *Act respecting the Barreau du Québec* explicitly acknowledges the overlap between the practice areas of *avocats* and those of *notaires* in non-contentious, private law situations.[71] When discharging the same duties as a *notaire*, however, an *avocat* is not subject to the rules regarding impartiality and public interest. In this sense, Québec *avocats* who practice as transactional lawyers have professional ethical expectations similar to those of English solicitors. Although the *Civil Code of Québec* provides that certain legal instruments must be executed by a *notaire*,[72] most clients may choose to visit a partisan *avocat* or an impartial *notaire* based on their personal needs and preferences.[73]

e. Adversarial vs. Inquisitorial Ethics

Although Québec's *avocat-notaire* separation offers valuable lessons to the construction of a restructured legal profession, one must acknowledge the inquisitorial ethics that underlie Québec's regulatory regime. In Québec, both *avocats* and *notaires* assume a secondary position in the administration of justice, which is led primarily by judges. In fact, Québec law expects *avocats* and *notaires* to "*collabore[r] à l'administration de la justice*."[74] *Larousse* defines *collaborer* as "*Travailler de concert avec quelqu'un d'autre, l'aider dans ses fonctions ; participer avec un ou plusieurs autres à une œuvre commune*."[75] As such, Québec's legal system defines *avocats* and *notaires* as collaborators in the justice process, reflecting the inquisitorial system's emphasis on delivering substantive justice over partisan adversaries.

Furthermore, Québec *notaires* are strictly regulated as holders of public office. The idea that *notaires* are agents of the State also shows influence from the civil law tradition. Therefore, copying and pasting Québec's *notaire* system into common law North America would lead to issues of incompatibility between the adversarial values underpinning the common law and the heavy civil law influence guiding Québec's regulatory regime for lawyers.

2. Professional Parameters and Ethics of Juridical Scriveners

a. Areas of Practice

---

[70] Québec, Code of Professional Conduct of Lawyers, s. 23.
[71] Québec, Act respecting the Barreau du Québec, s. 129(*e*).
[72] Québec, Notaries Act, s. 15(1).
[73] In addition to different professional ethics, *notaires* do not act *on behalf of* a client since they are public officers. Clients with complex legal situations that may require representation must retain an *avocat*. See Éducaloi, "Notaire."
[74] Although the English version of the statutes provides that an *avocat* "participates in" and a *notaire* "takes part in" in the administration of justice, the French version use the identical expression of "*collabore à*." Supposing the legislature means for the same expression to express the same idea, one should look at the word "*collaborer*" to interpret the meaning of the statutes.
[75] "Working together with and helping someone else in their duties; participate with others in a common work" (translation by author).

In my proposed legal profession, the right of appearance before any adjudicator is reserved for Advocates. This arrangement enables an Advocate to be a "resolute warrior of their client in the purest form." Advocates, and *only* Advocates, can practice in adversarial settings because the professional ethics of Advocates are tailored to the adversarial system. Naturally, Juridical Scriveners focus their practice on non-adversarial situations (e.g., drafting of conveyances, general legal advice, notarization). Just like current transactional lawyers, Juridical Scriveners are able to issue letters on behalf of their clients and are free to explore alternative dispute resolution with their clients.[76] Unlike current transactional lawyers, however, Juridical Scriveners are under the *obligation* to refer their client to an Advocate once a file is court-bound, similar to how English solicitors must refer their clients to a barrister when a trial is inevitable. Unlike in the English system, however, clients in the proposed system need not keep both the Scrivener and Advocate on retainer once the litigation commences.

Besides clearly adversarial settings like trials and non-adversarial settings like commissioning an affidavit, there is a gray area of contentious legal affairs like negotiation. I propose that this gray area remains the exclusive practice area of Juridical Scriveners. One of the main fail-safes in the adversarial system is the existence of an impartial adjudicator. Generally speaking, the adjudicator's decision can be further reviewed for correctness. This feature of the adversarial system mitigates the risks of putting Advocates in a position where they have to defend the distasteful positions of their clients. This safety valve, however, does not exist in contentious situations like negotiation or mediation. Even though a third party may exist, they do not have the same level of authority and reviewability as arbiters and judges. Therefore, this type of case is best left to Juridical Scriveners who are not governed by pure adversarial ethics. Furthermore, ADR processes (other than arbitration) require the drafting of mutually agreeable terms of resolution, an area that falls conveniently into Juridical Scriveners' expertise.

b. Professional Ethics

Since Juridical Scriveners act without the check of an adjudicator (and without pushbacks from opposing counsel in non-contentious cases), a Scrivener must not hide behind adversarial ethics and supposed "ignorance" of the social and moral consequences of their professional conduct. Since the nature of the work undertaken by Juridical Scriveners lays the foundation of consequent legal relationships and subsequent litigation, Scriveners cannot be morally ambivalent to the content of their work. The new regulatory framework will loosen the requirements that Juridical Scriveners must take or retain a client *vis-à-vis* the moral values of the client's cause. Under the new scheme, Juridical Scriveners will be personally invested (not *fully* accountable since the client is the one making orders) in the moral nature of the service they render. The new scheme will also ask Juridical Scriveners to step out of their "lawyer hat" and engage in candid moral conversation with their clients.

i. Moral Investment in Client's Cause

[76] Although arbitration is part of the general concept of ADR, it involves an impartial third party issuing binding decisions, thus falling into the exclusive competency of Advocates as they appear "on behalf of a party to the legal matter before an adjudicator."

Juridical Scriveners hold an informational advantage compared to their clients—the practitioner knows the law in that area and likely has experience dealing with similar situations. The informational asymmetry makes Juridical Scriveners not mere technicians but collaborators in achieving the goals of their clients. This is especially true in areas of law known to be convoluted, even for those with legal training, such as land law, estate law, and tax law. Unlike an individual trying to take advantage of their cash-stripped friend with a plain-language contract stipulating near-usury interest rates, those who seek to conceal their assets from their spouses, creditors, or the government are incapable of achieving anything resembling their goal without the help of a lawyer specializing in that field of law.

As Kennedy points out, "the better your legal skills, the less neutral you become."[77] If a Juridical Scrivener is extremely proficient at creating shell entities to conceal income and assets, their retainer with a cartel or a multi-billion-dollar company is less of a kitchen knife being used in a murder but more of a sniper rifle being used in an assassination. The same logic applies when the government seeks legal opinion on justifying torture or when a corporation seeks to legitimize its falsified operations.[78]

As such, a Juridical Scrivener is morally invested—not necessarily accountable, but at the bare minimum invested—in the professional service they render. The law routinely denies the façade of agnosticism to "inadvertent" drug traffickers[79] and those who "unknowingly" breach fiduciary duty,[80] accusing them of "willfully blindness." Lawyers, however, are rarely even *willfully blind* since clients are straightforward with their undesirable goals. Therefore, Juridical Scriveners must have some moral nexus to the fruits of their intellectual labor.

ii. Interests of Third Parties and Society

Since Judicial Scriveners are at least somehow aiding clients in achieving their causes, the practitioners ought to consider the implications of their work on people other than their clients. In adversarial processes, interested third parties and the State (the Crown in Canada) may choose to intervene, or a judge may appoint an *amicus* to lend a voice to the unheard stakeholders. In non-adversarial ones, however, no one other than Judicial Scriveners is in the position to bring in the voices absent from the table. This is true even in negotiations where an opposing party might be represented since terms of conflict resolution may have an indirect impact on third parties. The duty to consider third-party and societal interests is primarily procedural. It is a Judicial Scrivener's job to brief their client on the potential consequences of a given course of action, and this includes potential harm to third parties and to society.

The consideration for third-party and societal interests is not futile even when the client might have assessed the external impacts of their cause before retaining a Scrivener's service. Since the client's specialized knowledge tends to be in different fields, adding the Scrivener's analysis only complements the existing calculations undertaken by the client. However, Judicial Scriveners are private practitioners and not

---

[77] Kennedy, "The Responsibility of Lawyers for the Justice of Their Causes," 1160.
[78] Vischer, "Legal Advice as Moral Perspective."
[79] R. v. Lagace, 2003 CanLII 30886 (ON CA) (Court of Appeal for Ontario 2003); R. v. Farmer, 2014 ONCA 823 (CanLII) (Ontario Court of Appeal 2014).
[80] Y.R.C.C. NO. 890 v. RPS Resource Property Services, 2010 ONSC 3371 (CanLII) (Superior Court of Justice 2011).

public officers like Québec *notaires*. As such, they are not forbidden from continuing with any given course of action upon completing an analysis of social consequences.

iii. Scrivener-Client Candor

At common law, the communications between a person and their counsel are entitled to privilege. This protection enables clients to speak to their lawyer in utmost confidence, helping the lawyer devise legal strategies and advice for their client. This candor, however, should be a two-way street. While the client is expected to be candid and open about unfavorable facts and history, the Juridical Scrivener should be candid and honest about unfavorable opinions they may have formed against the client's cause.

This candor serves multiple purposes. The first is that it informs the client of a potential conflict of interest. Although legal training enables practitioners to wholeheartedly advocate for stances that would be personally unpalatable, the client may very well desire a lawyer who *actually* believes in their cause. In this sense, being candid with the client about one's moral assessment of the situation is a liberating exercise *both* for the Scrivener and for the client.

Secondly, just like with a lawyer's assessment of third-party impacts, a Scrivener's moral perspective may be novel to a client, thus contributing to a conducive conversation about the ethical implications of the client's cause. Just like a normative discussion with a friend, however, it befalls upon neither the Scrivener nor the client a duty to convince. Even if a Scrivener and a client decide to "agree to disagree," they have engaged in a meaningful exploration of moral possibilities, thus enhancing the mutual trust they hold in each other.

iv. Conclusion

Developing a moral investment in a client's cause is not meant to *forbid* or *dissuade* Scriveners from representing clients with distasteful causes. Instead, the new scheme asks Juridical Scriveners to account for their decision to represent a particular client in a particular way. If a lawyer has just been admitted to the Order of Juridical Scriveners and is drowning in student debt, they are free to take on affluent clients with distasteful causes so long as the practitioner is willing to admit "I chose my clients according to their ability to pay."[81] Similarly, if a lawyer attempted to discuss the moral implications of a client's cause with the client to no avail, all that moral investment asks of the Scrivener is an *honest* "I have tried my best."

Although sounding perfunctory, this process of morally "coming clean" puts the fraction of transactional lawyers habitually evading moral responsibility outside their comfort zone. Now that the excuse of being bound by the "cab rank rule" to be a "zealous advocate" is rightfully unavailable, transactional lawyers need to re-examine the relationship between their work and their morals. It is worth emphasizing that it is always an individual's responsibility to assess whether their conscience is clean, and the new professional ethics regime only seeks to prevent lawyers from avoiding the moral question.

---

[81] Kennedy, "The Responsibility of Lawyers for the Justice of Their Causes," 1160.

IV. FUTURE RESEARCH

A. In-House Advisors

The last segment of my proposed legal profession is In-House Advisors. Under the current system, in-house counsel have multiple roles in a firm: legal consultant, counsel, and employee.[82] Since they have to advise the firm and may need to represent the firm in court, their duties are an aggregate of that of an Advocate and of a Juridical Scrivener. Furthermore, as an in-house lawyer is an employee, the nature of their relationship with the client is different than that of an Advocate and of a Scrivener.[83] The dominant model, however, refuses to acknowledge this set of unique circumstances facing in-house lawyers. Indeed, Lord Denning held that in-house lawyers "are regarded by the law as, in every respect, in the same position as those who practice on their own account."[84] As such, a new regulatory scheme must account for the conditions of in-house counsel and set appropriate boundaries for their professional ethics.

B. Compensation Scheme for Advocates

When I proposed the Advocate-Scrivenor-Advisor trichotomy, I suggested that all Advocates be employed by a single employer, who charges clients based on their means and pays Advocates on a predictable and merit-based scale. The financial viability of this scheme, however, warrants further investigation. Future research can borrow from Legal Aid programs and the current compensation scheme for medical doctors in Canada to refine the proposal for a centralized employer.

CONCLUSION

The dominant model of professionalism is rightfully critiqued for its rampant misuse by certain transactional lawyers to escape ethical responsibility for their clients' causes. This paper Advocates for a new set of professional ethics for the segment of the legal profession called Juridical Scriveners, who are limited to processing transactional accounts. The proposed scheme requires Scriveners to make a personal moral investment in the services rendered to clients. It is not the purpose of this proposed scheme, nor of any model of sustainable professionalism that values the different developmental needs of individual legal practitioners,[85] to annihilate the possibility of lawyers practicing for questionable clients with the *sole purpose* of profiting from the arrangement. However, it is important to transform the role of lawyers in non-adversarial settings from that of an agnostic laborer to a morally concerned holder of knowledge. This added moral perspective contributes to candid and honest

---

[82] Shari L. Klevens and Alanna Clair, "Ethical Considerations for In-House Lawyers," *Dentons* (blog), October 5, 2023, https://www.dentons.com/en/insights/newsletters/2023/october/5/practice-tips-for-lawyers/ethical-considerations-for-in-house-lawyers.

[83] Gregory Richards, "Encountering and Responding to Ethical Dilemmas and Professional Challenges in the Role of In-House Counsel," *WeirFoulds LLP* (blog), April 14, 2009, https://www.weirfoulds.com/encountering-and-responding-to-ethical-dilemmas-and-professional-challenges-in-the-role-of-in-house-counsel.

[84] Ken B. Mills, "Privilege and the In-House Counsel," *Alberta Law Review* 41, no. 1 (2003): 79; Alfred Crompton Amusement Machines v. Customs and Excise Commissioners (No. 2).

[85] Farrow, "Sustainable Professionalism," 98-100.

communication between lawyers and clients while enhancing clients' autonomy by explicitly addressing any moral disagreements between the lawyer and the client.

BIBLIOGRAPHY

1. Alfred Crompton Amusement Machines v. Customs and Excise Commissioners (No. 2) (House of Lords).

2. Bar Standards Board. The BSB Handbook (2023). https://www.barstandardsboard.org.uk/for-barristers/bsb-handbook-and-code-guidance/the-bsb-handbook.html.

3. Brooks, Neil. "The Judge and the Adversary System." In *The Civil Litigation Process Cases and Materials*, edited by Janet Walker, Garry Watson, Timothy Pinos, and Allan C. Hutchinson, 7th ed., 89–119. Toronto: Emond Montgomery Publications, 2010.

4. Éducaloi. "Notaire." Éducaloi. Accessed January 27, 2024. https://educaloi.qc.ca/capsules/notaire/.

5. Farrow, Trevor C. W. "Sustainable Professionalism." *Osgoode Hall Law Journal* 46, no. 1 (January 1, 2008): 51–103. https://doi.org/10.60082/2817-5069.1207.

6. Felstiner, William L.F., Richard L. Abel, and Austin Sarat. "The Emergence and Transformation of Disputes: Naming, Blaming, Claiming…." *Law & Society Review* 15, no. 3/4 (1980): 631. https://doi.org/10.2307/3053505.

7. Groia v. Law Society of Upper Canada, 2018 SCC 27. Accessed December 1, 2023.

8. Hutchinson, Allan C. "Fighting Fair - A Call to Ethical Arms." *The Professional Lawyer* 23, no. 3 (2016): 12–16.

9. Kennedy, Duncan. "The Responsibility of Lawyers for the Justice of Their Causes," 1987. https://hdl.handle.net/2346/87006.

10. Klevens, Shari L., and Alanna Clair. "Ethical Considerations for In-House Lawyers." *Dentons* (blog), October 5, 2023. https://www.dentons.com/en/insights/newsletters/2023/october/5/practice-tips-for-lawyers/ethical-considerations-for-in-house-lawyers.

11. Mills, Ken B. "Privilege and the In-House Counsel." *Alberta Law Review* 41, no. 1 (2003): 79.

12. Ontario. Law Society Act, R.S.O. 1990 c. L.8.

13. Pye Tait Consulting and Bar Standards Board. "Provision of Legal Services by Barristers," May 2017. https://www.barstandardsboard.org.uk/asset/C551BDF1-4C2E-404C-B7A11B802F31A2C3/.

14. Québec. Act respecting the Barreau du Québec, CQLR c. B-1.

15. ———. Code of ethics of notaries, CQLR c. N-3, r. 2.

16. ———. Code of Professional Conduct of Lawyers, CQLR c. B-1, r. 3.1.

17. ———. Notaries Act, CQLR c. N-3.

18. R. v. Farmer, 2014 ONCA 823 (CanLII) (Ontario Court of Appeal 2014).

19. R. v. Lagace, 2003 CanLII 30886 (ON CA) (Court of Appeal for Ontario 2003).

20. Richards, Gregory. "Encountering and Responding to Ethical Dilemmas and Professional Challenges in the Role of In-House Counsel." *WeirFoulds LLP* (blog), April 14, 2009. https://www.weirfoulds.com/encountering-and-responding-to-ethical-dilemmas-and-professional-challenges-in-the-role-of-in-house-counsel.

21. Solicitors Regulation Authority. SRA Standards and Regulations (2019). https://www.sra.org.uk/solicitors/standards-regulations/.

22. The Bar Council. "Direct Access Portal." The Bar Council. Accessed January 29, 2024. https://www.barcouncil.org.uk/bar-council-services/for-the-public/direct-access-portal.html.

23. Vischer, Robert K. "Legal Advice as Moral Perspective." SSRN Scholarly Paper. Rochester, NY, August 10, 2005. https://papers.ssrn.com/abstract=771006.

24. Y.R.C.C. NO. 890 v. RPS Resource Property Services, 2010 ONSC 3371 (CanLII) (Superior Court of Justice 2011).

THREE ISSUES OF DATA COMPLIANCE GOVERNANCE IN CHINA BASED ON CASE ANALYSIS

Teng Zhang*

Abstract: With the rapid development of China's digital economy and increasingly stringent global data protection regulations, corporate data security and compliance have become key factors that determine the competitiveness and sustainable development of enterprises. However, China's data compliance governance still faces three main challenges: inadequate self-regulation of corporate data compliance, imperfect internal management system, and the risk of violations throughout the entire data lifecycle. To address these issues, China needs to comprehensively improve data security and compliance of Chinese enterprises by improving the system to strengthen corporate behavior guidance, intensifying internal compliance mechanisms, and identifying the compliance process for data processing. Based on the world's most typical legal systems related to data security and corporate compliance, and drawing on local Chinese judicial cases, this paper provides an in-depth analysis of the current challenges faced by Chinese enterprises in data security compliance. Finally, the paper explores and proposes a comprehensive governance path. Through the implementation of the comprehensive measures proposed in this study, Chinese enterprises now have practical guidance in the increasingly complex data security environment, thereby promoting their stable and sustainable development.

Keywords: Data Compliance; Chinese Enterprises; Corporate Governance; Information Security; Compliance

* School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, China.

Table of Contents

Introduction 239

I. Related Work 240

II. Overview of Data Security Legislation in China, Europe and America 241

III. Data Compliance Trends Among Chinese Enterprises 245

A. The Issue of Data Circulation Security is Severe 246

B. The Legal Risks of Data Leakage and Information Regulation Have Increased 248

C. The Governance Dilemma of Data Ownership Confirmation and Data Flow 248

IV. CHALLENGES OF DATA COMPLIANCE IN CHINA 249

A. Inadequate Self-Regulation of Corporate Data Compliance 249

B. Imperfect Internal Management System 250

C. The Risk of Violations Throughout the Entire Data Lifecycle 251

V. Exploration of the Compliance Path for Data Security in Chinese Enterprises 252

A. Improving the System to Strengthen Enterprise Behavior Orientation 252

B. Intensifying the Internal Compliance Mechanism 253

C. Identify the Compliance Process of Data Processing 253

Conclusion 254

INTRODUCTION

In recent years, China has issued numerous laws and regulations focusing on data security and compliance governance, placing increasing emphasis on the issue of compliance governance for data security in enterprises. Corporate compliance refers to proactive prevention of the risks of civil sanctions, administrative penalties, and reputation damage by complying with laws and regulations.[1] Data compliance encompasses the key aspects of traditional corporate compliance and introduces new requirements for the handling of corporate data: firstly, it requires adherence to rules and regulations to ensure that business activities are not punished; secondly, it mandates the establishment of a sound data compliance management system to prevent risks such as data infringement and data leakage.

However, according to the "White Paper on Enterprise Data Compliance (2021)" issued by the China Software Testing Center[2], most enterprises closely related to data and digital economy face issues such as failure to fulfill their related compliance obligations in data handling, lack of a well-established compliance system, and escalating difficulties in security governance. These problems expose the vast amount of data stored by enterprises to significant risks of data leakage, and the social impact caused by these risks is also showing an expanding trend.

With the increasing requirements for data security and compliance, enterprises are facing unprecedented challenges and opportunities. They must establish an effective compliance governance system based on compliance with laws and regulations to ensure sustainable business development. Data compliance not only focuses on the traditional physical security of data but also on the content security of the information contained in the data. Therefore, strengthening the legality and compliance of enterprise data security has become a significant topic in current business operations, academic research, market regulation, and other scenarios.

Based on a comparison of data compliance laws between China, Europe, and the United States, as well as an analysis of current judicial cases in China, there are still three major issues in China: inadequate self-regulation of corporate data compliance, imperfect internal management system, and the risk of violations throughout the entire data lifecycle.

Data compliance cannot be achieved overnight. To address the existing related issues, we should approach from three perspectives: first, improving the system to strengthen corporate behavior guidance, and guiding enterprises to establish a suitable compliance system based on their own characteristics and needs from a legal perspective; second, enhancing the internal compliance mechanism of enterprises and building a management framework for data compliance from within the enterprise; third, identifying the compliance process for data processing to ensure that legal risks can be eliminated throughout the data lifecycle.

---

[1] Zhang, Y. H. (2019). Criminal compliance: International trends and Chinese practices. *Procuratorial Daily*. November 2, p. 3.

[2] China Software Testing Center. (2022). Official website. Retrieved from https://www.cstc.org.cn/search.jsp?wbtreeid=1001

Considering the increasing attention paid by legal academics and legal practitioners to data compliance issues in China, data compliance research remains scarce and one-sided. This paper delves into three aspects: the current state, challenges, and coping strategies for data compliance in China. It aims to promote research on Chinese data compliance in the global legal community and address current issues of data security compliance governance in Chinese enterprises.

I. RELATED WORK

Currently, academic research on enterprise data compliance focuses primarily on the introduction and implementation of a criminal compliance system for data security, legal regulations on data circulation and cross-border data management, and the construction and optimization of a data protection compliance system.

Li Bencan (2018)[3], Yu Chong (2020)[4], and Zhang Yong (2022)[5] believe that it is necessary to implement a criminal compliance system to address the increasing complexity and frequency of data crimes. Through co-governance between enterprises and the state, they aim to strengthen data security protection, achieve active general prevention, and promote the convergence of criminal and administrative measures to achieve collaborative governance.

Wang Liming (2023)[6], Paul de Hert et al. (2016)[7], and Xu Duoqi (2020)[8] argue that in the era of the data economy, enterprise data circulation and cross-border data management face challenges, emphasizing the need to promote compliance, facilitate circulation, and ensure security through improved laws, regulatory mechanisms, and technical means.

Chen Ruihua (2020)[9], Mao Yixiao (2022)[10], and He Hang (2022)[11] believe that in the process of digital transformation, enterprises should establish and optimize a data protection compliance system, achieving a shift from passive compliance response to proactive compliance governance. This ensures data security and compliance, thus strengthening the data governance structure and compliance management processes to address the challenges and risks of data protection compliance.

Based on the aforementioned research context, current scholars' research on enterprise data compliance still ignores the effectiveness of industry self-regulation

---

[3] Li, B. (2018). *Compliance and Criminal Law: A Global Perspective*. China University of Political Science and Law Press.

[4] Yu, C. (2020). The iterative alienation of data security crimes and the path of criminal law regulation: From the perspective of the introduction of criminal compliance programs. *Journal of Northwest University (Philosophy and Social Sciences Edition)*, 5, 93-102.

[5] Zhang, Y. (2022). The crime filtering model of criminal compliance in data security. *Academic Forum*, 3, 13-24.

[6] Wang, L. (2023). Data Protection by Civil Law. *Digital Law*, (1), 43-56.

[7] Paul & Vagelis. (2016). The new General Data Protection Regulation: Still a sound system for the protection of individuals? *Computer Law & Security Review*, 32(2), 179-194.

[8] Xu, D. (2020). On the legal guarantee of two-way compliance for cross-border data flow regulation enterprises. *Eastern Law Review*, (2), 185-197.

[9] Chen, R. (2020). Basic issues of corporate compliance. *China Legal Review*, (1), 178-196.

[10] Mao, Y. (2022). Research on data protection compliance system. *Journal of the National Procurators' College*, (2), 84-100.

[11] He, H. (2022). Key issues and relief measures of corporate data security compliance governance. *Guizhou Social Sciences*, (10), 126-133.

systems, the perfection of internal management systems, and the widespread existence of data violation risks in enterprises. Therefore, this paper focuses on three prominent issues: insufficient self-regulation systems for enterprise data compliance, imperfections in the internal management system, and the risk of violations throughout the entire data lifecycle. To address these issues and ensure the effective protection and legitimate utilization of electronic data, as well as to maintain enterprise security and data-driven operations, the author believes that the problems of data security compliance governance in Chinese enterprises today should be solved by improving systems to strengthen enterprise behavior orientation, intensifying internal compliance mechanisms, and clarifying the compliance process for data processing.

II. OVERVIEW OF DATA SECURITY LEGISLATION IN CHINA, EUROPE AND AMERICA

As humanity enters the digital age, the severity and possible spillover effects of data security risks have had negative impacts in multiple fields such as politics, technology, economy, and society. Therefore, the scientific response to data security is bound to become a key focus of research in the digital age. People's attention has shifted from the possibilities and extensibility of "Being Digital"[12] to the characteristics and forms of the "Network Society"[13], and further to topics such as data security and artificial intelligence in the "Cyber Society"[14]. As the network society and the real society become increasingly integrated, security issues in cyberspace are becoming more prominent. Computer crimes have also evolved into cybercrime, manifesting in three basic types: cybercrime where the network serves as the "criminal object", "criminal tool", and "criminal space"[15]. However, regardless of the type of cybercrime, its manifestations always revolve around electronic data and information technology composed of "0" and "1" in "bits".

With the digital transformation of economic and social life, people have realized that data is a fundamental strategic element of the digital economy. Data has become the foundation for individuals to participate in economic and social activities, for businesses to carry out operational activities, and for social public organizations and state agencies to fulfill their duties. The protection of data security has become an important guarantee for the healthy development of the digital economy.

Currently, legislation in Europe, the United States, and other countries is mainly divided into two categories: data jurisdiction and data protection.

Firstly, there is the legislation targeting data jurisdiction in the United States. Leveraging the authority granted by the Foreign Corrupt Practices Act (FCPA), the US government utilizes its formidable national strength and global influence. Through joint law enforcement actions between the Department of Justice and the Securities and Exchange Commission, it actively mobilizes various departments such as the Federal Bureau of Investigation, Homeland Security, and the Criminal Investigation Division of the Internal Revenue Service. It employs diverse legal means including wiretaps,

---

[12] Negroponte, N. (1997). *Being Digital* . Hainan Publishing House. (Original work published 1995)
[13] Castells, M. (2010). *The Rise of the Network Society*. Blackwell Publishing. pp. 500-509.
[14] Zheng, Z., & He, M. (2004). Analysis and discrimination of the concept of "network society." *Sociological Research*, 1, 9.
[15] Yu, Z. (2014). Intergenerational evolution of cybercrime and the response of criminal legislation and theory. *Qinghai Social Sciences*, 2, 1-11.

undercover operations, search warrants, and subpoenas, and has established a whistleblower hotline and reward mechanism. These efforts are fully coordinated to combat overseas corruption and maintain a fair and just international business environment. When overseas businesses conducting operations in the US or having business relationships with US companies are suspected of violations, they often seek lenient criminal and administrative penalties through agreements and the establishment of a compliance governance system. However, many companies still pay heavy fines[16]. With the advent of the information age, the US has formulated a cybersecurity strategy to clarify national interests and responsibilities[17]. The US government has expanded its "long-arm jurisdiction" experience under the FCPA to protect personal information, privacy, and data through domestic legislation at different levels. As long as there is a jurisdictional connection with the US, companies from anywhere must accept the jurisdiction of US law. Although this model has been opposed by many countries since its inception, accusing the US of violating the basic principles of international law with its "long-arm jurisdiction," most have to accept this reality due to the US's absolute advantages in technology, finance, politics, military, and soft power.

The second aspect is data protection legislation in EU countries, which originated from the Western society's strong emphasis on privacy protection. As the information age surged in the 1990s, the risk of privacy breaches increased daily. To address this situation, the Privacy Impact Assessment (PIA) system emerged as a response and gradually became a conventional measure to protect individual privacy. The EU provides high-standard guarantees for data security by establishing a sound and rigorous rule system. In 1995, the "Directive on the Processing of Personal Data and Data Protection" conducted a preliminary exploration of the operational mechanism of data protection agencies in EU countries in the form of binding governance rules for the first time. With the continuous improvement of Internet technology and the advent of the big data era, the legal environment in the EU has undergone profound changes. In May 2018, the General Data Protection Regulation (GDPR) came into effect, regarded as the strictest personal information protection and data regulation[18]. Based on the Privacy Impact Assessment (PIA), the EU has specially established a Data Protection Impact Assessment (DPIA). As the core content of the EU data protection framework, DPIA has not only become a key approach to personal data security governance in various countries but also provides a path guide for corporate compliance management, economic digital transformation, and the construction of a data security governance legal system in various countries.

Since China joined the international Internet system in 1994, the internet has gone through nearly 30 years of ups and downs in China. The focus has shifted from an initial emphasis on computer and cybercrime, to network security protection at the national security level, and finally to addressing the challenges of personal information protection and data security risks in the information society. The cyberspace governance legislative process has shown a three-phase incremental development trend, including legislation on infrastructure and legal relationships, legislation on network

---

[16] Yang, K., & Tao, D. (2017). The US Foreign Corrupt Practices Act. *China Economic Weekly*, 49, 2.
[17] Yu, L. (2012). The impact of American internet strategy on China's political and cultural security. *International Forum*, (2), 7.
[18] Jiao, N. (2022). Research on the operational mechanism of data protection agencies in EU countries. *Information Magazine*, 41(5), 154-161.

information services and industry management, and legislation on information security, data security, and online transactions. Specifically:

The first aspect is legislation on infrastructure and legal regulatory systems. In this phase, administrative legislation was adopted to address the issues of the foundation for internet development and network security from the perspective of network infrastructure. Regulations and rules related to the network infrastructure and basic behaviors were successively formulated.

Table 1. *Data Security Guarantee and Network Transaction Legislation.*

Year	Effectiveness Level	Laws and Regulations
2013	Administrative Regulations	Regulations on the Protection of Computer Software
2011	Administrative Regulations	Regulations on the Security Protection of Computer Information Systems
1997	Departmental Rules	Interim Provisions on the Administration of International Networking of Computer Information Networks

The second aspect is legislation on the Internet information services and industry management systems. In 2013, Edward Snowden, a former employee of the Central Intelligence Agency (CIA) and the National Security Agency (NSA) in the United States, exposed the "Prism" program[19]. This program revealed the serious harm caused by the United States' use of the Internet to conduct network surveillance and cyberattacks on other countries, which endangered the network security and information security of various countries. Based on this, Chinese legislative bodies began targeted and systematic legislative activities to address the issue of cyberspace governance and issued the following laws and regulations.

Table 2. *Data Security Guarantee and Network Transaction Legislation.*

Year	Effectiveness Level	Laws and Regulations
2011	Administrative Regulations	Measures for the Administration of Internet Information Services
2013	Administrative Regulations	Regulations on the Protection of Information Network Transmission Rights
2013	Departmental Rules	Regulations on the Protection of Personal Information of

[19] Zou, Q. "Prism Gate" causes huge legal controversy in the US. Retrieved from https://www.chinacourt.org/article/detail/2013/06/id/1014622.shtml

		Telecommunications and Internet Users
2017	Departmental Rules	Measures for the Administration of Internet Domain Names
2017	Departmental Rules	Regulations on the Administration of Internet News Information Services
2019	Departmental Rules	Regulations on the Administration of Blockchain Information Services"
2019	Departmental Rules	Regulations on the Administration of Financial Information Services
2011	Administrative Regulations	Measures for the Administration of Internet Information Services
2013	Administrative Regulations	Regulations on the Protection of Information Network Transmission Rights

The third aspect is data security guarantee and improvement of the network transaction law. This stage focuses mainly on strengthening and improving cyberspace governance through the enactment of basic laws by the legislature. Based on maintaining national security and data sovereignty and competing for dominance in cyberspace and data flow governance with developed European and American countries, the following laws and regulations have been formulated and implemented:

Table 3. *Data Security Guarantee and Network Transaction Legislation.*

Year	Effectiveness Level	Laws and Regulations
2015	Law	National Security Law
2017	Law	Cybersecurity Law
2019	Law	E-commerce Law
2020	Law	Cryptography Law
2021	Law	Data Security Law
2021	Law	Personal Information Protection Law
2021	Administrative Regulations	Regulations on the Safety Protection of Critical Information Infrastructure
2022	Departmental Rules	Measures for the Safety Assessment of Data Export

2022	Departmental Rules	Measures for Cybersecurity Review

With the promulgation of the Data Security Law in June 2021, the data governance system achieved a breakthrough from scratch, becoming a fundamental law in the field of security in China. According to the provisions of China's Data Security Law, data security refers to the state of effective protection and lawful utilization of data through the adoption of necessary measures, as well as the ability to maintain a continuous state of security. The Data Security Law also clearly states that the protection of data security aims to "protect national sovereignty, security, and development interests" and to "promote data development and utilization, and protect the legitimate rights and interests of citizens and organizations."

III. DATA COMPLIANCE TRENDS AMONG CHINESE ENTERPRISES

As mentioned earlier, corporate compliance refers to conforming to legal requirements.[20] The legal norms in this region can be broadly categorized into four main types: Firstly, the national legal system, which encompasses laws, administrative regulations, administrative rules, local regulations, and judicial interpretations. All normative documents with legal authority are binding guidelines that must be followed. Secondly, business practices, which include both written norms such as codes of conduct issued by various industry associations, as well as unwritten business practices and ethics. Thirdly, internally developed company rules and regulations. Violations of rules established by the company itself may also result in penalties for the business. Lastly, international organizational treaties, such as compliance management and sanctions systems established by international organizations such as the World Bank. For violations of treaty obligations by companies, the World Bank can impose sanctions with conditions for lifting, setting a probationary period of several years for the company and requiring it to rebuild its compliance program.

Based on the research conducted by Chinese scholars, corporate compliance, as a form of corporate governance, encompasses three aspects: administrative compliance, criminal compliance, and compliance with international sanctions. Specifically:

Firstly, in terms of criminal compliance. The criminal compliance mechanism views corporate compliance as a key basis for prosecution, conviction, and sentencing, and also serves as an important basis and content for corporate settlement agreements. This mechanism has been established as a statutory basis for corporate criminal liability and is presented as an alternative, informal criminal procedure in criminal procedure law.[21] Therefore, criminal compliance refers to a legal system in which prosecutors set a probationary period for corporate compliance when a company is suspected of specific crimes, urging the company to make a commitment to compliance and actively implement rectification measures. After inspection and evaluation by a third-party organization, lenient treatment is implemented for the relevant companies based on the actual situation.[22]

---

[20] Zhang, Y. H. (2019). Criminal compliance: International trends and Chinese practices. *Procuratorial Daily*. November 2, p. 3.

[21] Chen, R. H. (2020). Basic issues of corporate compliance. *China Legal Review*, (1), 178-196.

[22] Wang, L., & Wang, J. (2022). Active Duty to Enhance the Quality and Efficiency of Corporate Compliance Governance. *Prosecutorial Daily*, April 11, p. 3.

Secondly, in terms of administrative compliance, it refers to a legal system in which administrative regulatory agencies use specific institutional designs and regulatory measures to ensure that companies comply with administrative laws and regulations, prompting them to meet compliance standards. If a company suspected of violating administrative laws has established a compliance system, it can be used as a legal basis to obtain more lenient or flexible administrative treatment.

Thirdly, anti-international sanctions compliance refers to an international legal system where international organizations that establish compliance management agencies require companies that violate the trading rules set by the organizations to establish effective compliance programs in exchange for the lifting of sanctions and penalties.[23]

Whether from the perspective of the enterprise's own development, the long-term stability and security of society, or the strategic layout of the country, corporate compliance is the inevitable path for the vigorous development of enterprises and even the national economy. With the development of the digital economy, the importance of data security has not only become the key to economic stability and development but also a focus of national security concerns. As the main participants in the digital economy, enterprises' compliance governance in the field of data security is the foundation and guarantee for resolving data security risks. Generally speaking, it encompasses two levels of meaning: one is to ensure the security of the data itself, and the other is to meet the security requirements of the environment where the data resides[24]. Specifically, compliance governance in the field of data security requires that enterprises' behaviors during operation should comply with relevant national laws, regulations, and rules related to data security, and must not violate the internal rules and regulations formulated by the enterprise.

A. The Issue of Data Circulation Security is Severe

In the era of digital economy, economic globalization is entering a new era led by data flows. Data is increasingly becoming the core driving force for the vigorous development of economic activities, and compliance processing of data flows has become increasingly important. Economic globalization has enabled businesses to operate beyond geographical and national boundaries, while the rapid development of the Internet has further narrowed the distance between people around the world. However, in the process of carrying out cross-border operations or financing, enterprises will inevitably face the issue of cross-border data flows.

Cross-border data flows have become increasingly common. According to the *White Paper on Compliance and Technical Application of Cross-border Data Flows* issued by the Cross-border Data Flows Group of the Open Islands Open Source Community[25], China has emerged as a leading country in digital trade and is actively accumulating and utilizing massive amounts of data. Based on predictions from International Data Corporation (IDC), the average annual growth rate of data volume

[23] Chen, R. (2022). *Basic Theory of Corporate Compliance*. Legal Publishing House.
[24] Chen, B., & Hu, Z. (2021). The legal path to coordinate data security and development in the digital economy. *Changbaixuekan*.
[25] Digital Elite Network. (2023). Retrieved from https://www.digitalelite.cn/h-nd-5751.html

in China is expected to reach approximately 30 percent from 2021 to 2025, making China a leader in global data volume.

In the future, the digitization of the global economy driven by data elements will bring broader and more profound impacts. The acceleration of cross-border data flows is changing the global economic landscape. In the long run, this will affect the competitive relationship between developed and developing countries in the labor market, reshape the global labor market, and further influence the global industrial chain layout and value chain division of labor. Therefore, regions and enterprises with advantages in data elements and intelligent technology are expected to occupy the mid-to-high-end position in the global value chain.[26]

The cross-border flow of data elements will have profound and widespread impacts on core areas such as international benefit distribution, national and cyber security, and data sovereignty. In view of this, we must maintain a high level of awareness and vigilance regarding the security issues involved in data circulation. In 2021, the incident of Didi Chuxing's listing in the United States[27] triggered widespread concern about the security risks of "cross-border data circulation" across the internet. As a direct result, the National Internet Information Office of China imposed administrative penalties on Didi Chuxing, including stopping the registration of new users, removing the app from all platforms, issuing a notice, conducting a cybersecurity review, and imposing a fine of RMB 8.026 billion on Didi Global Inc.[28] The main violations committed by Didi Chuxing include illegally collecting and abusing personal information data, engaging in data processing activities that pose a serious threat to national security, and illegal operational behavior, which bring serious security risks to the security of national critical information infrastructure and data security. On December 18, 2020, the Chinese stock Luckin Coffee, which was listed in the United States, admitted to financial fraud[29]. This triggered the formal effectiveness of the U.S. Congress's legislation on the supervision of listed companies, the "Holding Foreign Companies Accountable Act", which empowers the U.S. Securities and Exchange Commission (SEC) to initiate stronger information disclosure requirements for Chinese stocks listed in the United States in order to protect investors' interests. It requires Chinese stocks to hand over audit working papers, otherwise, it will suspend trading or delist Chinese stocks that do not submit audit working papers for three consecutive years. Although this incident, which has continued to this day, was temporarily suspended from the delisting crisis of Chinese stocks after negotiations between China and the United States reached a cooperation agreement on August 26, 2022. However, the cross-border audit review issue related to audit working papers is undoubtedly closely related to data circulation security. The urgent risk of data circulation security brought about by the above typical cases once strengthened the Chinese government's

---

[26] Hong, Y., Zhang, M., & Liu, Y. (2022). Promoting the safe and orderly flow of cross-border data to lead the globalization of the digital economy. *Bulletin of the Chinese Academy of Sciences*, 37(10), 1418-1425.

[27] Cyberspace Administration of China. (2022). Official website. Retrieved from http://www.cac.gov.cn/2022-07/21/c_1660021534306352.htm.

[28] Wang, Z.,&Zhang, F. (2022). The Cyberspace Administration of China Imposes a Fine of 8.026 Billion Yuan on Didi, Cheng Wei and Liu Qing Each Fined 1 Million Yuan. People's Network. Retrieved from http://finance.people.com.cn/n1/2022/0721/c1004-32481985.html

[29] China Securities Regulatory Commission. (2020). Notification on the Investigation and Disposition of Financial Fraud in Ruixing Coffee. Retrieved from http://www.csrc.gov.cn/csrc/c100028/c1000725/content.shtml

legislative work on this issue. In July 2022, the Cyberspace Administration of China issued the "Measures for the Safety Assessment of Data Exit", and subsequently formulated and issued the "Guide for the Declaration of Safety Assessment of Data Exit (First Edition)", further building a comprehensive solution that focuses on institutional norms and integrates technological innovation in the field of cross-border data flow.

B. The Legal Risks of Data Leakage and Information Regulation Have Increased

With the widespread application and rapid development of digital technology, its utilization of personal information has become increasingly frequent and diversified. However, this also poses escalating risks to individual rights and interests. To mitigate these risks, it is particularly crucial for enterprises to achieve compliance in information data management. Relevant enterprises primarily serve individual users. According to Chinese laws and regulations on cyberspace governance, the collection of personal information under a real-name system is indispensable. During the generation, collection, and utilization of personal information data, the primary data security issues faced by enterprises include: firstly, the illegal and non-compliant collection and use of personal information, such as collecting sensitive personal information without the individual's knowledge or using personal information beyond the scope agreed upon by the data subject; secondly, the failure to adopt effective technical and management measures to protect data, leading to the theft, leakage, or damage of important data collected or generated during operations, thereby posing a threat to national security, public interest, and individual rights and interests.[30]

On the other hand, according to legal provisions, enterprises engaged in information services can be divided into three categories: network access services, information content services, and information service platform services. In the process of providing the aforementioned internet information services, enterprises are faced with obligations to supervise user behavior, information content, and enterprise qualifications. Meanwhile, if enterprises themselves are engaged in value-added telecommunications services, internet news information services, internet religious information services, internet live streaming services, and other fields, they still need to obtain corresponding licenses and registrations. Failure to obtain the corresponding licenses or registrations, or violations of relevant legal provisions resulting in serious consequences, may result in heavy administrative penalties or the revocation of licenses and registrations.

C. The Governance Dilemma of Data Ownership Confirmation and Data Flow

As an important component of assets, data possesses the attribute of eventually turning into capital through flow and transactions, similar to other assets.[31] The capitalization of data highlights the transformation of data's role, elevating it from being merely an information carrier to becoming an asset with the potential to create personal

---

[30] Feng, D., Zhang, M., & Li, H. (2014). Big data security and privacy protection. *Chinese Journal of Computers*, 37(1), 13.

[31] Lian, Y. (2021). *Data Rights Law 3.0: Legislative Prospects of Data Rights*. Social Sciences Academic Press.

or corporate wealth. This change has given the data market unprecedented and vast development opportunities.

In the course of business operations, enterprises produce and collect vast amounts of data. However, enterprises only possess de facto control over these data, which is the right to use the data, but this does not equate to ownership of the data. Therefore, in the face of the emerging data market with tremendous development potential, how to monetize the data at hand, turn it into capital, and create more wealth has become a key concern for enterprises.

The current ambiguous definition of data ownership in China has led to many difficulties in data sharing and openness, making it necessary to establish compliance for data handling. To achieve free flow, efficient allocation, and fair competition in the data element market, determining data ownership has become an urgent priority.[32] Coase Theorem suggests that "if transaction costs are zero, no matter how rights are defined, optimal allocation can be achieved through market transactions, independent of legal provisions."[33] Therefore, relying on market regulation, ambiguous data property rights and inappropriate ownership relations may negatively impact market operations. In the current legal environment where data ownership is not clearly defined, enterprises can only address market and legal risks in data circulation by establishing an effective compliance governance system that meets regulatory requirements.

IV. CHALLENGES OF DATA COMPLIANCE IN CHINA

A. Inadequate Self-Regulation of Corporate Data Compliance

In today's data-driven social landscape, self-regulation of industry data compliance affects the benchmark of corporate governance. However, there are significant deficiencies in China's self-regulatory system in the field of data compliance, which can be deeply analyzed from two levels: industry norms and technical capabilities.

Taking data scraping behavior as an example, from the perspective of industry norms, the lack of forward-looking legislation and ineffective implementation of norms have become the two main issues. Starting from the ten verdict cases related to data anti-unfair competition released by the Beijing Intellectual Property Court[34], it is observed that the court frequently applies the principled provisions of the Anti-Unfair Competition Law when dealing with data unfair competition cases. This indicates that there are gaps in the specific provisions of existing laws on data scraping behavior, and related data scraping behaviors often lack clear legal regulations.

The use and processing of data are constantly evolving, yet existing industry norms often struggle to keep pace with these changes, exhibiting a legislative lag. Industry norms lack specific regulations on data scraping, and in some industries, there

---

[32] Zhang, Y., & Zhang, B. (2022). Data ownership confirmation and institutional response in the context of building a data element market. *Journal of Shanghai University of Political Science and Law: Rule of Law Forum*, 37(4), 20.

[33] Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3(1), 1-44.

[34] "Top Ten Typical Cases of Data-Related Anti-Unfair Competition in Beijing Intellectual Property Court." Retrieved from Beijing Intellectual Property Court website: https://bjzcfy.bjcourt.gov.cn/article/detail/2023/07/id/7382298.shtml

are no norms whatsoever related to data collection. This has led to widespread data violations among enterprises. In terms of norm implementation, even when relevant industry norms exist, enterprises often only symbolically comply with them in practice, falling far short of the standards expected by the norms. Both of these major issues identified through case analysis illustrate the significant lack of enforcement in industry self-regulation, with enterprises generally lacking the intrinsic motivation to strictly adhere to the norms.

From the perspective of technical specifications, the current technical specifications may have been inadequately considered during their formulation, making it difficult to achieve comprehensive and effective supervision in practical applications. The "Information Technology Big Data Data Classification Guide"[35] is overly theoretical, and its data classification methods and standards are disconnected from actual operations. Even if enterprises read this specification, it is difficult for them to have a clear definition of data classification, making it challenging for the specification to play its due regulatory role in practical applications. When formulating technical specifications, if similar specifications are not adequately scientifically argued and practically tested, the scientific and practical value of the specifications will be greatly reduced. In this case, even if enterprises try their best to comply with these specifications, it may be difficult to achieve the desired data compliance effects.

B. Imperfect Internal Management System

As enterprises increasingly rely on data, it is particularly important to establish a comprehensive internal management system. Currently, many enterprises still have many deficiencies in this area, mainly reflected in the loopholes in their internal data management processes and the urgent need to improve compliance mechanisms.

The so-called loopholes in the enterprise's internal data management process mainly refer to institutional defects or inadequate management in storage, access control, and other links. Taking the case of Han Bing's destruction of a computer information system[36] as an example, as a database administrator, he should have shouldered the responsibility of protecting corporate data security. However, due to the imperfect internal management system, Han Bing used his privileges to conveniently log in to the financial system, deleted financial data and related applications, resulting in the company's financial system being completely inaccessible. This fully exposes the company's deficiencies in data classification management, permission settings, and data operation procedures, failing to effectively prevent internal personnel from abusing their privileges, which led to a serious data security incident.

Today, enterprises are facing severe challenges and a constantly changing legal environment in terms of data management and compliance, thus their compliance mechanisms urgently need to be improved. Enterprises need to establish clear compliance standards and operational procedures in various aspects such as data collection, storage, use, and destruction. Didi Global Inc. was once heavily fined for

---

[35] State Administration for Market Regulation, Standardization Administration. (2020). *Information technology - Big data - Guidelines for data categorization*. GB/T 38667-2020.
[36] Beijing First Intermediate People's Court. (2020). Criminal Verdict, (2020) Jing 01 Xing Zhong 490.

not establishing clear data compliance standards and operational procedures.[37] Currently, many enterprises are still struggling in this regard, lacking systematic and forward-looking compliance planning. The lack of this section often leaves enterprises at a loss when facing data compliance risks, making it difficult to effectively respond, and ultimately leading to penalties.

C. The Risk of Violations Throughout the Entire Data Lifecycle

The risk of violation, which runs throughout the entire life cycle of data, is an important issue that cannot be ignored by enterprises. In the 1980s, the concept of life cycle was transferred to the field of information management[38], referring to the process of information data from generation, utilization to elimination. To be closer to the current enterprise's actual application of data, the enterprise data life cycle can be divided into four stages: data collection stage, data storage and use stage, data sharing and trading stage, and data destruction stage.

In the data collection stage, enterprises collect a large amount of data, and the risks of violation mainly focus on three core aspects: first, the compliance of data collection methods, especially the use of crawler technology; second, the randomness of data collection, where disorderly or improper operations occur occasionally; third, the excessiveness of data collection, which means excessive collection may infringe privacy or violate relevant regulations.

The administrative penalty imposed by the China Market Supervision and Administration Bureau on Chengdu Zhike Technology Co., Ltd.[39] provides a thorough illustration of the administrative penalty risks arising from non-compliant data collection methods. The excessive and indiscriminate collection of data is further exemplified in the case of Shanghai Yuanyun Investment Management Co., Ltd. suspected of excessively collecting consumer personal information[40]. Both cases reflect the current compliance issues faced by enterprises in the data collection process, including unauthorized collection of personal and platform data, failure to clearly state the purpose, method, and scope of data collection, as well as indiscriminate and excessive collection methods. These issues significantly increase the risk of administrative penalties for enterprises, such as hiccups to business operations and damage to public welfare.

In the data storage process, the risk of violation mainly stems from neglecting the establishment of data security management systems and operating procedures, as well as failing to implement deidentification and encryption measures when storing

---

[37] Wang, Z.,&Zhang, F. (2022). The Cyberspace Administration of China Imposes a Fine of 8.026 Billion Yuan on Didi, Cheng Wei and Liu Qing Each Fined 1 Million Yuan. People's Network. Retrieved from http://finance.people.com.cn/n1/2022/0721/c1004-32481985.html.

[38] Levitan, K. B. (1982). Information resources as “goods” in the life cycle of information production. *Journal of the American Society for Information Science*, 33(1), 44-54.

[39] Chengdu Zhike Technology Co., Ltd. Improper Conduct Case [No. 51010023000187]. (2024). Chengdu Administration for Market Regulation.

[40] Shanghai Yuanyun Investment Management Co., Ltd. (2023). Suspected of excessively collecting consumers' personal information case. Shanghai Jing'an District Market Supervision and Administration Bureau. Case No. Hu Shi Jian Jing Chu [2023] 062021002963.

data. The Guangdong Driving Training Data Security Case[41] exemplifies this point. Most enterprises have significant shortcomings in data storage management, not only lacking a data backup mechanism but also failing to set a clear data storage period. In terms of handling stored personal information, enterprises also appear to be insufficiently rigorous. They use personal information directly after collection without necessary processing, exacerbated by the problems of privacy leakage and abuse.

The risk of violation during the data sharing and trading phase focuses on requiring the data receiver to fulfill the obligation of data security protection. In the country's first case of commercial secret infringement involving a data transaction buyer[42], San Company, as the data receiver, failed to fulfill the corresponding data security obligations and even disclosed and used the information to third parties. This shows that during data sharing and trading, if enterprises do not take any measures to ensure the obligations of the data receiver, it can easily lead to data misuse, deviating from the original purpose authorized by the data owner. In the current increasingly strict administrative law enforcement environment for personal information protection, these reckless behaviors of enterprises, if discovered, will inevitably face serious legal consequences.

During the data destruction phase, it is necessary to regularly destroy data to prevent the accumulation of data, which can trigger the illegal risks mentioned above in the storage, sharing, and trading phases. The Cyberspace Administration of China conducted a cybersecurity review on CNKI. Based on CNKI's deficiencies in the data destruction phase, such as not providing an account cancellation function and not deleting users' personal information in a timely manner after canceling their accounts, administrative penalties were imposed on CNKI[43]. Failure to fulfill the obligation to destroy data not only leads to penalties for the company itself but also significantly increases the probability of cybersecurity risks such as citizen data breaches, which in turn affects national security.

V. EXPLORATION OF THE COMPLIANCE PATH FOR DATA SECURITY IN CHINESE ENTERPRISES

A. Improving the System to Strengthen Enterprise Behavior Orientation

Based on an in-depth analysis of the aforementioned issues, the primary challenge faced by enterprises in building a compliance system is the lack of a clear institutional orientation. To address this problem, relevant laws, regulations, and reference plans should be actively introduced to guide enterprises in establishing suitable compliance governance systems based on their own characteristics and needs. In the future, China's legislative work on data security can start from two aspects: basic regulations and administrative rules and regulations. Legal regulations should be

---

[41] "Guangdong's First Case! A Company in Guangzhou Was Punished by the Police for Not Fulfilling the Obligation to Protect Data Security." (2022) Guangdong Legal Website. Retrieved from https://www.gdzf.org.cn/yasf/content/post_123060.html.

[42] Chongqing Guang Motorcycle Manufacturing Co., Ltd. v. Guangzhou San Motorcycle Co., Ltd. (2022). Infringement of business secrets dispute. Chongqing Free Trade Zone People's Court. Verdict No. (2022) Yu 0192 Min Chu 8589.

[43] Cyberspace Administration of China. (2023). The Cyberspace Administration of China Imposes Administrative Penalties Related to Cybersecurity Review on CNKI According to Law. Retrieved from https://www.cac.gov.cn/2023-09/06/c_1695654024248502.htm

imposed on key aspects such as data ownership confirmation, circulation, and transaction security.

In terms of basic regulations, legislation needs to clarify the ownership, use rights, and management rights of data, providing a solid legal foundation for data transactions. This includes detailed provisions on data classification, data ownership, and legitimate methods of data acquisition, to ensure the circulation of data under legal and secure premises.

In terms of administrative regulations and rules, the focus should be on regulating the data circulation market, establishing strict market access standards, clarifying the rules and procedures for data transactions, and increasing penalties for violations. This aims to promote the healthy development of the data market while ensuring data security, thus guaranteeing the continuous improvement of the industry.

B. Intensifying the Internal Compliance Mechanism

The deficiencies in the internal management system of enterprises have seriously affected their data security and compliance. In order to enhance the data management capabilities of enterprises and reduce potential risks, enterprises must immediately proceed to improve their internal management systems, clarify the responsibilities and authorities of various departments and personnel, and establish scientific and reasonable data processing procedures. At the same time, enterprises should also strengthen employee training and awareness efforts to ensure that every employee fully understands the importance of data security and compliance and puts it into practice.

Currently, the "Compliance Management Measures for Central Enterprises"[44] provide a template for state-owned enterprises to establish a sound compliance management system tailored to their actual situation, including compliance systems, improving operational mechanisms, cultivating a compliance culture, and strengthening supervision and accountability. Private enterprises should learn from this approach to strengthen their internal management system. Regarding data security and compliance issues, private enterprises should deeply understand and draw on the core philosophy and practical experience of the "Compliance Management Measures for Central Enterprises" to build and improve their own compliance management system.

C. Identify the Compliance Process of Data Processing

To ensure that data complies with existing processing procedures and to reduce data breaches and other security risks, enterprises must clarify the compliance process for data processing. The "Information Security Technology - Security Capability Requirements for Big Data Services"[45] provides regulations for every aspect of the entire data processing chain, serving as an important reference for enterprises in handling data today. This standard not only provides clear guidance for data collection,

---

[44] State-owned Assets Supervision and Administration Commission of the State Council. (2022, August 23). Measures for the Compliance Management of Central Enterprises. Decree No. 42 of the State-owned Assets Supervision and Administration Commission of the State Council. Effective from October 1, 2022.

[45] National Information Security Standardization Technical Committee. (2023). Information security technology: Security capability requirements for big data service (GB/T 35274-2023).

storage, processing, and transmission but also emphasizes the importance of key security measures such as data encryption, access control, and security auditing.

To strictly implement the data classification and grading system, enterprises should follow the "Data Security Technology - Data Classification and Grading Rules"[46] for standardized classification and grading of data after collecting it. These rules help enterprises reasonably classify data grades based on sensitivity, importance, and value, thereby implementing more refined data management and protection measures.

Based on data classification and grading, enterprises also need to establish detailed data processing procedures and security policies to clarify access rights, encryption measures, backup strategies, and emergency response plans for each level of data. By implementing these measures, enterprises can ensure that data is fully protected during transmission, storage, and processing, effectively preventing the risk of data leakage and misuse.

After improving the aforementioned processes, enterprises should also regularly review and update their data security management systems and procedures to adapt to the constantly changing security threats and technological environment. Continuous enhancement of data security training and awareness raising is necessary to ensure that employees in the corresponding sectors understand and follow data security regulations, collectively maintaining the data security of the enterprise.

CONCLUSION

In the context of the rapid development of the digital economy, Chinese enterprises are facing increasingly severe challenges in data security and compliance. This study conducted an in-depth exploration of the core issues related to data security and corporate compliance among Chinese enterprises, revealing critical problems such as inadequate self-regulation of corporate data compliance, imperfect internal management system, and the risk of violations throughout the entire data lifecycle. To address these challenges, this paper systematically explores effective data security and compliance governance pathways, including improving the system to strengthen corporate behavior guidance, intensifying internal compliance mechanisms, and identifying the compliance process for data processing. These measures aim to ensure that enterprises fully comply with laws and regulations when processing data, thereby effectively guaranteeing the security and compliance of corporate data. This study not only provides strong theoretical support and practical guidance for Chinese enterprises in data security and compliance governance but also offers some reference for addressing corporate data security and compliance issues globally.

[46] National Information Security Standardization Technical Committee. (2024). Data security technology: Rules for data classification and grading (GB/T 43697-2024).

This page intentionally left blank.



www.ijlet.org

La Nouvelle Jeunesse

ISSN 2769-7142

9 772769 714009