

“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, USA.

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 <i>Especially</i> When They Don't Understand the AI's Process	79
B.	What Are the Dangers of Embracing this New Techno-Judiciary?	81
		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 atrophy 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 of 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 of 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%20see%20Maximillian%20Schreiner,%20Deepmind%20Chinchilla:%20Artificial%20Intelligence%20is%20far%20from%20being%20fed%20up/](https://the-decoder.com/gpt-4-has-a-trillion-parameters/#:~:text=Further%20details%20on%20GPT%2D4’s,Mixture%20of%20Experts%20(MoE).But%20see%20Maximillian%20Schreiner,%20Deepmind%20Chinchilla:%20Artificial%20Intelligence%20is%20far%20from%20being%20fed%20up/) (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\(MoE\).](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 Times 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 Castelvecchi, *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 Variability 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) SYNTHÈSE 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 Castelvecchi, *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 atrophy 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.