Legal Priorities
Project

Value alignment for advanced artificial judicial intelligence

Christoph Winter
Nicholas Hollman
David Manheim

LPP WORKING PAPER N° 5-2022

In: *American Philosophical Quarterly* (2023) 60 (2): 187–203.
doi.org/10.5406/21521123.60.2.06

American Philosophical Quarterly
Volume 60, Number 2, April 2023

VALUE ALIGNMENT FOR ADVANCED ARTIFICIAL JUDICIAL INTELLIGENCE

Christoph Winter, Nicholas Hollman, and David Manheim

ABSTRACT This paper considers challenges resulting from the use of advanced artificial judicial intelligence (AAJI). We argue that these challenges should be considered through the lens of value alignment. Instead of discussing why specific goals and values, such as fairness and nondiscrimination, ought to be implemented, we consider the question of how AAJI can be aligned with goals and values more generally, in order to be reliably integrated into legal and judicial systems. This value alignment framing draws on AI safety and alignment literature to introduce two otherwise neglected considerations for AAJI safety: specification and assurance. We outline diverse research directions and suggest the adoption of assurance and specification mechanisms as the use of AI in the judiciary progresses. While we focus on specification and assurance to illustrate the value of the AI safety and alignment literature, we encourage researchers in law and philosophy to consider what other lessons may be drawn.

KEYWORDS law & artificial intelligence; advanced artificial judicial intelligence (AAJI); value alignment; specification; assurance

1. Introduction

Judges serve complex functions and roles within society (Sourdin and Zariski 2013). They need to align their decisions with laws and policies (Epstein, Landes, and Posner 2013), and with their social functions, such as to settle disputes, resolve unclarity and conflicts in the law, and protect the rule of law, among others (Green 2016). Delegating judicial decision-making to artificial intelligence (AI) likewise necessitates value alignment. That is, decisions made and actions taken by an AI in this context ought to capture the complex norms, ideals, and broader goals of the judiciary. While previous work has focused on which abstract values judicial decision-making ought to be aligned with, such as principles of nondiscrimination (e.g., Pasquale 2015; Chen 2019; Kleinberg, Ludwig, et al. 2018), and transparency (e.g., Berman 2018; Coglianese and Lehr 2019; Winter 2022), this paper focuses on the challenges of alignment itself: How can we correctly specify and ultimately encode more concrete values into judicial AI in a way that ensures that these values are followed or respected as intended? What lessons can be drawn from existing research on alignment?[1]

Despite the concerns shared with the field of AI safety, the judicial AI literature has not yet recognized the value alignment framing and thus neglects issues and research directions that are discussed in research on the alignment problem more generally (see Christian 2020; Hendrycks et al. 2022; Hubinger et al. 2019; Kenton et al. 2021; Langosco et al. 2022; Leike et al. 2018; Manheim and Garrabrant 2019; Ngo 2022a; Russell 2019; Soares 2018). For instance, the AI safety literature shows many examples of an AI behaving in unintended ways due to a misaligned goal of the system (Amodei and Clark 2016; Amodei et al. 2016; Lehman et al. 2020; Krakovna 2018; "Artificial Intelligence Incident Database" 2022). Amodei and Clark illustrate this with an example of an incident where a game-playing AI trained to run a race course ended up maximizing its score by hitting targets without ever finishing the race, thereby meeting the specified goal of a high score while failing to satisfy the intended, but unspecified, outcome of completing the course.[2] By considering and learning from such past mistakes, technical and legal researchers alike can mitigate and avoid future bad or unintended outcomes in AI decisions[3]—particularly in more consequential settings, like the judiciary. Since the alignment literature proposes concrete methods to avoid misalignment, we argue that the alignment framing and its associated literature have much to offer the ongoing discussion. This may be especially important now, as AI becomes an integral part of the judiciary and begins to carry out judicial functions.

Currently, AI is used in judiciaries around the world to support or replace parts of human judicial decision-making (cf. Sourdin 2021; Winter 2022). *Supportive AI* provides information, advice, or other support to human actors. Risk assessment tools are a common form of supportive AI, used in criminal justice to aid a human judge with decisions about pretrial release and sentencing by predicting risk and likelihood of recidivism (Becker and Ferrari 2020; HLR 2017; see also Hayward and Maas 2021, pp. 13–14). For example, COMPAS, a risk assessment tool used in a number of states in the US, produces sentencing recommendations based on an interview with the offender and information from their criminal history (HLR 2017). Other jurisdictions use similar tools, such as HART in the UK and VICTOR in Brazil (Jauhar et al. 2021). In contrast, *replacement AI* carries out judicial decision-making, thereby leaving human actors in an even more limited role or replacing them entirely. Replacement AI does not merely support human actors by providing specific information or suggestions. Instead, this form of AI actually decides matters. While AI has not yet replaced judicial decision-making by humans, some applications have come close. For instance, China's Internet Courts use machine learning technology to generate judgments and decrees for judges in internet-related disputes (Beijing Internet Court 2019, p. 10; Shi, Sourdin, and Li 2021, p. 11; Stern et al. 2021).

Although current AI still has limited capabilities and responsibilities in judicial decision-making, economic incentives favor increased (and more widespread) use of AI to automate judicial decision-making. Re and Solow-Niederman (2019, p. 258) argue that "the pitch to invest in 'better, faster, cheaper' justice will prove irresistible" as courts find themselves overburdened with cases, which strains access to justice.[4] Furthermore, recent survey data suggests that, on average, legal scholars estimate that almost 30 percent of judicial decision-making will be carried out by AI in only 25 years time (by 2046), tripling their estimate that AI currently accounts for less than 10 percent of judicial decision-making (Martínez and Winter 2022). In other surveys, AI experts[5] estimated that there is a 50 percent chance that AI that surpasses human performance in every task,[6] including decision-making in any domain, will be developed by 2059 (Stein-Perlman, Weinstein-Raun, and Grace 2022), 2060 (Zhang et al. 2022: cross-sectional sample), 2061 (Grace et al. 2018; Walsh 2018), and 2076 (Zhang et al. 2022: panel sample).[7]

Scholars have considered the possibility of advanced AI in the judiciary, referring to it as "cyborg judges" (Crootof 2019), "AI judges" (Michaels 2020), "global digital arbiter" (Maas 2022), and the "legal singularity" (Alarie 2016; Deakin and Markou 2020). In this paper, we focus on one specific manifestation of replacement AI: advanced artificial judicial intelligence (AAJI), defined as "an artificially intelligent system that matches or surpasses human decision-making in all domains relevant to judicial decision-making" (Winter 2022). Survey data suggests that legal academics by-and-large consider the use of AI in the judiciary a promising development in the long-run, especially from the perspective of improving access to justice, economic efficiency, predictability and even judicial independence and transparency (Martínez and Winter 2022); however, scholars also note that this transition could threaten constitutional rights (Huq 2020b), the rule of law (Greenstein 2021; Zalnieriute, Moses, and Williams 2019), judicial independence (Sourdin 2021; Winter 2022), and basic judicial legitimacy (Crootof 2019; Michaels 2020; Sourdin 2021; Winter 2022).

This article frames the transition to AAJI as a domain-specific value alignment problem. Instead of discussing the merits of specific goals and values, such as fairness or non-discrimination, we consider the question of *how* AAJI might be aligned with those goals and values to be reliably integrated into legal and judicial systems over time. This parallels a broader discussion about how transformative AI must be aligned with human values generally. The paper proceeds as follows. In Section 2, we introduce the value alignment problem for AAJI. We then discuss two areas of AI safety and corresponding insights for judicial values and the design of AAJI, with Section 3 on specification and Section 4 on assurance. Section 5 concludes.

2. Value Alignment

For the use of artificial intelligence in the judiciary to be beneficial to humans, judicial AI must necessarily be aligned with core judicial values. The value alignment problem, then, is the challenge of ensuring that AI acts according to these values (see also Hilton 2022, fn. 29 reviewing various definitions of alignment). Researchers have described various forms of alignment, such as behavior alignment (Leike et al. 2018), intent alignment (Christiano 2018), incentive alignment (Everitt et al. 2021a; Everitt et al. 2021b), outer and inner alignment (Hubinger et al. 2019), as well as various approaches to solving it, such as iterated amplification (Christiano et al. 2018), debate (Irving et al. 2018), and recursive reward modeling (Leike et al. 2018; for general overviews, see Everitt et al. 2018; Kenton et al. 2021). While recognizing the imprecise nature of the terms "AI alignment" and "value alignment" (cf. Christiano 2018; Christian 2020; Gabriel 2020), we nevertheless find it helpful to consider value alignment for AAJI as being primarily concerned with two challenges: (1) the *normative challenge* of what values or objectives an AI ought to apply in the judicial setting and how, and (2) the *technical challenge* of how to encode such values or objectives into an AI (cf. Gabriel 2020; Ngo 2020). Although these are described as two distinct challenges, they are interrelated in various ways, as we describe in later sections on specification and assurance.[8]

Notwithstanding the novelty of the value alignment framing for AI in the judiciary, legal scholars have long considered this first, normative challenge, of what fundamental values in the judicial system describe the main functions and responsibilities of the judiciary. Most notably with regards to judicial values in the United States, Martínez and Tobia (2022) find that the majority of US law professors consider fairness, economic efficiency, welfare, and rule-of-law values to be among the primary purposes that law should serve.[9] Further important values identified in the literature include judicial independence (cf. T. S. Ellis III 2005; Shetreet 2012), transparency (Samaha 2008), and predictability (Bednar 2004). Others have described the roles and values of the judiciary in terms of how they are shaped. Garoupa and Ginsburg (2009) discuss the role of internal audiences (other judges) and external audiences (lawyers, media, the public) in maintaining accountability and a good reputation for the individual judge and the judiciary as a whole. Likewise, Banks and O'Brien (2021) describe a variety of internal and external constraints on judicial power and decision-making (see also Baum 2009; Spiller and Gely 2008).

While many values have also been discussed in the context of judicial AI, some judicial values remain especially neglected. Values that are discussed frequently pertaining to AI in the judiciary include transparency (Berman 2018; Coglianese and Lehr 2019; IEEE 2019; Sunstein 2019; Winter 2022; Zalnieriute, Moses, and Williams 2019), non-discrimination and impartiality (CEPEJ 2018; Chander 2017 reviewing Pasquale 2015; Chen 2019; European Commission 2021; Hacker 2018; Kleinberg, Ludwig, et al. 2018; Sourdin 2018; Sunstein 2019; Surden 2020; Winter 2022), and reason-giving in judicial decision-making (Atkinson, Bench-Capon, and Bollegala 2020; Huq 2020a; Pasquale 2019). However, few discuss the role an advanced judicial AI could have in the broader political landscape, such as its impact on democratic legitimacy. Notable exceptions include recent discussion on separation of powers and judicial independence (Sourdin 2021; Stern et al. 2021; Winter 2022), as well as the educational role that the judiciary, and judicial commentary in particular, arguably has towards litigants, lawyers, and civic education broadly in society (Sourdin 2018).

In the following, we discuss two specific contributions that the alignment framing and its growing literature can make to the discourse on judicial AI, in particular in adequately understanding both the problems posted, and potential solutions. We first argue that the current literature on values in the judiciary lacks adequate *specification* needed to build safe AAJI which raises issues on the theoretical level and for practical decision-making. More specifically, the literature fails to resolve tensions both within and between values, and neglects certain values. Second, we describe how *assurance*, a mechanism for control and accountability, is necessary for AAJI alignment, and argue that the existing discussion on accountability around AI legal decision-making should build on similar themes found in the AI safety and alignment literature. While we focus on specification and assurance to illustrate the value of the AI safety and alignment literature, we encourage researchers in law and philosophy to consider what other lessons may be drawn, for instance regarding robustness (Hendrycks et al. 2022, pp. 3–4; Kohli et al. 2019; Ortega, Maini, and DeepMind safety team 2018), how to address various forms of uncertainty (Christian 2020; Kanal and Lemmer 1985), and the importance of interpretability (for an overview, see Hubinger 2022).

3. Specification

Specification is the task of conveying to a machine learning system what exactly its designers would like it to do (Rudner and Toner 2021). Misspecification, then, is the gap between the ideal specification (what the designer intends the system to do) and the design specification (what objective the designer actually implements) or the revealed specification (what the system actually does to achieve the objective).[10] For some tasks, such as choosing which tiles in a CAPTCHA test contain a fire hydrant, it is relatively straightforward to write a precise description that avoids misspecification. Yet, for many other tasks, such as ensuring the right balance of judicial independence, access to justice, and rule-of-law values, it is difficult to capture the nuances of these values in sufficiently specified mathematical language that can be encoded into an AI. While specification is normative in that it entails describing (and therefore choosing) the values themselves, encoding those values in a sufficiently action-guiding way is part of the technical challenge. From this foundation, we argue for the importance of specification to resolve tensions *between* competing values and resolve tensions that arise from conflicting meanings *within* a given value (cf. Whittlestone et al. 2019) in the face of value tradeoffs that judges inevitably face. Finally, we discuss the importance of specifying values that consider the role of an AAJI in the broader political context, involving interactions with other branches of government and the public, in order to align with the entirety of the judicial role.

We turn first to tensions that arise *between* conflicting judicial values. To illustrate this, Kleinberg, Ludwig, et al. (2018) describe tension between values raised by the use of screening algorithms that predict flight and crime risk for criminal defendants pending trial. The algorithms in question use data on age, criminal history, and current offense to analyze the risks more accurately than human judges (see also Kleinberg, Lakkaraju, et al. 2018). By their nature, these algorithms can quantify tradeoffs between different values, in this case related to racial disparities in criminal justice (see generally SCPI 2017; Sentencing Project 2000), detention and incarceration rates, and public safety. They can enhance decision-making, but they also require an explicit, and perhaps uncomfortable (cf. Tetlock 2003), choice about competing values and outcomes which are less transparent in human-centered judiciaries.[11] For example, an algorithm with greater predictive power could reduce crime while maintaining the same detention rate or vice versa; it could be used without race as a data point, or it could seek greater racial balance in release rates (Sunstein 2019). Even though it has been noted that the algorithm developed by Kleinberg, Lakkaraju, et al. is capable of achieving greater racial equality while simultaneously reducing crime and detention rates (Sunstein 2019; Winter 2022), as long as racial equality and the reduction of crime and detention rates are not positively correlated, such trade-offs will have to be made eventually. By focusing on specification, the alignment literature urges one to think about how these uncomfortable trade-offs between competing values should be made to avoid the risks of under-specified systems.

Specification must also resolve tensions *within* values, that is, competing notions of the same value that would lead to different decisions or outcomes. For instance, terms such as “fairness” and “justice” can mask differences in interpretation across different normative perspectives. While fairness is generally accepted as a core value of the judiciary, the term is arguably under-specified in legal scholarship. Meanwhile, literature on algorithmic decision-making identifies several possible conceptions of fairness that may be sufficiently specified, yet some are incompatible with one another (see Berk et al. 2021; Binns 2018; Clouser and Gert 1990; Kleinberg et al. 2017; Ruf and Detyniecki 2021). Other widely endorsed judicial values for AI, such as the rule of law and democratic principles, face similar issues. How can we implement rule-of-law values when multiple definitions and interpretations thereof exist (see Berman 2018; Pasquale and Cashwell 2018; Tamanaha 2004)? Likewise, how can a specification describe the value of protecting human rights (see CEPEJ 2018; IEEE 2019; Wachter, Mittelstadt, and Floridi 2017; Završnik 2020) when multiple conceptions and interpretations of human rights exist and conflict in theory and in practice (Waldron 1989; Xu and Wilson 2006)? These questions remain neglected, yet are crucial for reaching a sufficiently safe specification of judicial AI, even in current applications. Encouraging the discussion to adopt a focus on tensions will be a useful tool moving forward, especially given legal and philosophical scholars' familiarity with balancing and proportionality tests in weighing competing values (Wilson 1995; but see Aleinikoff 1987; Winter 2020).

Furthermore, because human judges play an important role in a broader political context, the ideal specification of the values and objectives of an AAJI must account for checks and balances and public legitimacy, among other considerations. However, these values are rarely discussed in the context of judicial AI[12], and it is unclear how AAJI might fulfill them in practice. For example, judicial review in the United States is well-understood as the ability of the judiciary to strike down acts by the legislative or executive branch as unconstitutional, but the exact nature, scope, and ultimate value of this ability are highly contested (see, e.g., Grey 1975; Landau 2017; Waldron 2006). How can an AAJI system effectively maintain checks on power? How can values that seem almost too broad by definition, such as separation of powers and judicial independence, be sufficiently specified to avoid misspecification, which might result in unintended shifts of political power?

Through this value alignment lens, there is a clear need for more discussion on how the roles and values of the judiciary should be specified. Only through sufficient specification can an AAJI navigate tensions within and between values that arise in judicial decision-making and avoid the pitfalls of misspecification. As specification should also include values on the role of the judiciary in the broader political context, future research may rise to the challenge by combining insights from the alignment and AI safety literature with the long-lasting debates in political and constitutional theory.

4. Assurance

Once the values for an AAJI have been specified and encoded, it is important to monitor an AI's decision-making over time in a process known as assurance. Assurance is the collection of methods and mechanisms used to inspect, evaluate, and control AI and ensure that it operates in the intended way (Kazim and Koshiyama 2020; Ortega, Maini, and DeepMind safety team 2018). This section gives an overview of the assurance mechanisms developed in the AI safety literature and shows how they can aid AAJI value alignment. While assurance mechanisms take various forms, they typically aim at increasing the *verifiability*, *transparency & interpretability*, and *interruptibility* of AI (see also systematic review by Batarseh, Freeman, and Huang 2021).[13] We will address each in turn.

First is *verifiability*, or mechanisms for verifying claims about AI development and deployment, including as they relate to value alignment. For AAJI development, verification mechanisms could be used to evaluate how AI protects privacy or judicial independence and to ensure that AI developers follow applicable laws and policies. Brundage et al. (2020) survey a number of specific mechanisms at the software, hardware, and institutional level that could be used for verification purposes. Institutional mechanisms, which shape the incentives of those involved with AI development, may be especially valuable for legal-philosophical scholarship to explore, given the possibility for legal mechanisms to shape relevant incentives and researchers' familiarity with the subject. For example, regulations might require third party audits to verify claims made at different stages of AAJI development, such as those about sufficiently representative training data or decision outcomes, as well as ongoing audits throughout its use to ensure that the system is still aligned. Beyond specific mechanisms and incentives, there is much for scholars of law, policy and philosophy to unpack regarding the desired normative benchmarks for verification. What normative measures could allow third parties to verify sufficient judicial independence and access to justice? More broadly, what approaches can be used to verify different judicial values? What legal incentives for verification are possible?

Second, the AI assurance literature emphasizes the importance of *transparency* and *interpretability*, which refer to the ability to observe and understand how an AI works.[14] The importance of interpretability is highlighted by Deeks (2020), who describes the benefits of judicial reason-giving for internal and external audiences as improving decision quality, promoting efficiency, strengthening legitimacy, constraining other decision-makers, and fostering accountability (see also Coglianese and Lehr 2019; Christiansen and Eskridge 2014; Garoupa and Ginsburg 2009; Selbst and Barocas 2018). Transparency and interpretability were most notably at issue in *State v. Loomis*, where the defendant argued that the court's reliance on the COMPAS software program violated his due process rights in part due to lack of transparency. Because the methodology of the software is a trade secret, neither the defendant nor court could evaluate how the risk scores were determined or factors weighed (HLR 2017). Despite this, the court ruled that the COMPAS assessment could be used along with other information in making a decision. However, if political and philosophical ideals already favor strong interpretability norms and legal systems often require judges to offer explanations for their decisions (Surden 2020), one might wonder what if anything new there is to be gained from the alignment literature.

The literature on AI assurance can make valuable contributions to this debate. It promotes models that are interpretable by design over merely explainable AI, particularly for high-stakes decisions, including criminal justice (Rudin 2019; see also Rudin and Radin 2019). Crucially, it also motivates us to consider which interpretability methods, such as mechanistic interpretability for understanding neural language models (Olah 2022; Olah et al. 2020) may be most promising in the context of judicial AI. Furthermore, the alignment literature offers numerous methods for evaluating interpretability for different tasks and requirements that are applicable to judicial decision-making. More concretely, Doshi-Velez and Kim (2017) lay out three approaches for evaluating interpretability: application-grounded, human-grounded, and functionally-grounded, each with different costs and levels of specificity. Application-grounded approaches involve experiments in which human experts explain how an AI conducts a real-world task, such as deciding a case, and the quality of those explanations are evaluated to determine interpretability. Human-grounded approaches are also based on human experiments but with simplified tasks, such as having human evaluators choose between explanations offered for an AI's decision or behavior. Last, Doshi-Velez and Kim outline functionally-grounded evaluations that require an AI to meet a formal definition of interpretability that has been validated through application- or human-grounded approaches, thereby avoiding the need for additional human experiments. Given the importance of interpretability in judicial processes and the high stakes of judicial decisions, evaluation of AAJI should include at least some application-grounded approaches, even if resource-intensive. That said, further research could explore whether different approaches to interpretability are appropriate at different stages of development or for certain judicial functions or values. For example, human-grounded approaches might be appropriate for evaluating more general notions, such as having a lay audience choose which of two decisions is more understandable. Application-grounded evaluations may involve different experts for different types of tasks or values. What approaches to interpretability might be preferable for different functions and roles of the judiciary, and under what conditions? What kinds of tasks should be evaluated and how? What might be required by law in different jurisdictions? Could the right to a fair trial require high standards of interpretability?

Just as there are several methods for evaluating interpretability, the alignment literature also offers different kinds of interpretability that promise insights for AAJI assurance. Weller (2019) and Doshi-Velez and Kim (2017) distinguish between *global* interpretability, a general understanding of how the system works, and *local* interpretability, knowing the reasons for a specific decision. In the context of AAJI, global interpretability would be an understanding of patterns in judicial decision-making, while local interpretability would be knowing the reasons for a specific judicial decision. Crucially, both local and global interpretability could relate to questions of law and fact. The distinction between local and global interpretability may also be relevant to discussions of AAJI alignment, as different values may depend on different forms of interpretability. Deeks (2019) discusses how approaches to explain AI in the judiciary may pursue local or global interpretability, depending on the application, noting that global interpretability may be important to evaluate bias or error in the system overall, whereas local interpretability may be useful for an individual navigating a particular case. Although further investigation is necessary, global interpretability is likely necessary to uphold values that relate to the role of the judiciary in the broader political system, such as judicial independence and separation of powers. More generally, one might ask in what circumstances local or global interpretability are appropriate for AAJI, and under what conditions this might depend on the relevant audience (Deeks 2019; Garoupa and Ginsburg 2009).

Finally, assurance also entails the ability to *interrupt* an AI in order to stop or otherwise change undesired behavior.[15] Users and developers may want to interrupt an AAJI system for a variety of reasons, for example if it displays unintended behavior, relies on insufficiently representative training data, or undermines legal norms.[16] In the human-centered judiciary, these concerns are often addressed via an appeals process. Equally, assurance processes in an AAJI ought to account for the appeals process, but they may also need to allow immediate interruption in the event of a technical error or unintended behavior by the system—even more so when judicial AI moves from supportive to replacement roles. This raises a number of questions, such as what the requirements for interruption might be, which agents would have the power to interrupt, and how judicial independence might be upheld in such scenarios. Further, one might investigate whether the appeals process should be adapted to address new concerns from AAJI, for example to stay a judgment or order while seeking immediate review for misaligned behavior.

Because of the fundamentally different nature of AI and the opportunity for new failure modes, it is vital that an AAJI achieves a level of assurance that is similar or superior to that of human judges in order to support important judicial values. The discussion on AI in the judiciary may benefit greatly from advances and insights in the AI safety and alignment literature on assurance that have yet to be incorporated.

5. Conclusion

Jurisdictions around the world increasingly rely on AI in the judiciary. This article has argued that many of the challenges that accompany this transition and have yet to be addressed are alignment problems. Instead of discussing the merits of specific goals and values, we considered the question of *how* AAJI can be aligned with desired goals and values. More precisely, we argued that the alignment framing helps to identify two central questions: (1) how should judicial values be specified to provide practical guidance in concrete situations, including when they conflict (*specification*), and (2) what methods and mechanisms are needed to provide assurance that judicial AI operates on the values we intend (*assurance*)?

In order to appreciate the challenge of alignment for AAJI and answer these questions comprehensively, we suggested various strategies and research directions for scholars of law, philosophy, and computer science to pursue. We found that the alignment framing sheds light on the importance of specification that provides guidance for tensions between and within values like separation of powers and judicial independence. It also offers novel ways to approach assurance concerns, for instance with methods to verify values, evaluate interpretability, and ensure adequate judicial review and interruptibility. In all of these areas, the alignment literature provides indispensable insights for legal and philosophical scholarship on AAJI, and its framing offers a useful guiding function.

If AI is one day able to match or surpass human judicial decision-making, but that day is far enough in the future, we may be able to defer discussion. However, it is unclear how far off this future is, and it seems foolhardy to defer discussion until such systems are proven possible, since the proof is likely to come in the form of the actual use of these systems—in other words, too late to prepare.

Christoph Winter
Instituto Tecnológico Autónomo de México
Mexico City, Mexico /
Harvard University
Cambridge, MA, USA
Christoph_winter@fas.harvard.edu

Nicholas Hollman
Legal Priorities Project
Cambridge, MA, USA

David Manheim
Technion, Israel Institute of Technology
Haifa, Israel /
Foresight Institute
San Francisco, CA, USA

NOTES

For comments, discussion, and critique, we are grateful to Matthijs Maas, Noam Kolt, Jonas Schuett, Jesse Shulman, Laura Messner, Eric Martínez, Cullen O'Keefe, Markus Anderljung, Michael Aird, and the editors of the *American Philosophical Quarterly*. We are particularly grateful to Suzanne Van Arsdale for excellent research assistance and many useful suggestions.

[1] Conversely, the AI alignment literature has already begun to draw on legal scholarship, including in the fields of contract law and economics (see, e.g., Hadfield-Menell and Hadfield 2018) and human rights law (Bajgar and Horenovsky 2022).

[2] This type of misalignment has been referred to as "reward hacking" (see, e.g., Christiano 2016; Cohen, Hutter, and Osborne 2022; Irving et al. 2018; Leike et al. 2018; Skalse et al. 2022), which may become even more severe as AI systems become increasingly capable (Pan, Bhatia, and Steinhardt 2022). The literature on goal misgeneralization demonstrates a related type of misalignment, in which an AI system trained to pursue a specified objective nevertheless learns to pursue a proxy and thereby threatens AI safety (Shah et al. 2022).

[3] For example, several recommendations exist for mitigating the risk of reward hacking (see, e.g., Everitt et al. 2021; Zhuang and Hadfield-Menell 2020).

4. Wang (2020) explains how this takes on different forms in state-driven (China) and market-driven (US) adoption of AI in the judiciary.

5. The literature on forecasting indicates that domain expertise alone does not necessarily provide an advantage in making predictions (Tetlock 2005; Tetlock and Gardner 2015; Armstrong and Greene 2007, pp. 998–1002). For a review of past AI forecasts, see Muehlhauser (2016).

6. Also referred to as "Human-Level Machine Intelligence" (HLMI).

7. Note that the relevant survey questions differ slightly. Grace et al. 2018; Stein-Perlman, Weinstein-Raun, and Grace 2022; and Zhang et al. 2022 in their panel sample study asked about "unaided machines [that] can accomplish every task better and more cheaply than human workers" and instructed participants to "[i]gnore aspects of tasks for which being a human is intrinsically advantageous, e.g. being accepted as a jury member." These three surveys were further conditional on "human scientific activity continu[ing] without major negative disruption." Walsh (2018) asked about AI that "can carry out most human professions at least as well as a typical human," and Zhang et al. 2022 in their cross-sectional sample asked about AI "able to perform almost all tasks (>90% of all tasks) that are economically relevant better than the median human paid to do that task in 2019. You should ignore tasks that are legally or culturally restricted to humans, such as serving on a jury."

Karnofsky (2021a; 2021b) reviews a variety of forecasting methods, including expert predictions, about transformative AI development and raises the question of what the burden of proof should be when making predictions; in other words: "How good do these forecasting methods need to be in order for us to take them seriously?"

8. Gabriel (2020) questions the "simple thesis" that it is possible to work on the technical challenge separately from the normative challenge.

9. For a discussion regarding considerations of fairness and societal welfare in the legal system, see Kaplow and Shavell (2001). For an overview of the case for economic efficiency in judicial decision-making, see generally Posner (2003, p. 26).

10. For various ways in which misspecification might occur, see Kenton et al. (2021). A match between ideal and design specification has also been referred to as outer alignment, whereas inner alignment refers to a match between design and revealed specification. For an overview of the usage of these terminologies and framings, see Ngo (2022b).

11. Elsewhere, one of us (Winter 2022) refers to this new form of transparency as "transparency of options," arguing that judicial AI could increase the overall level of transparency in judicial decision-making, especially from a liberal democratic perspective. Note, however, that transparency should be distinguished from interpretability, which we turn to in the next section.

12. For notable exceptions, see Section 2.

13. Note that this list is not exhaustive, and not all mechanisms listed are useful exclusively for assurance purposes. For instance, interpretability might well be considered an integral part of specification.

14. These and other terms are used in the literature with varying definitions (cf. also Deeks 2019, pp. 1834–38). For example, Atkinson, Bench-Capon, and Bollegala 2020 use the term explainability in much the same way we use interpretability. That said, we prefer the approach by Kenton et al. (2021, p. 5), which differentiates between *explainability methods*, which are requested post-hoc of an output, and *interpretability methods*, which seek to give humans (*ex ante*) understanding of the internal workings of a system (see also Rudin 2019). The latter has also been referred to as the "decompositional approach" to explainable AI (Edwards and Veale 2017, p. 64). In response, Deeks (2019, p. 1835) termed the "exogenous approach" to explainable AI that provides information about how the model works using extrinsic, orthogonal methods rather than providing the reasoning or inner workings of the system.

15. The interruptibility of an AI further depends on it being "corrigible"—that is, it must cooperate with corrective intervention such as shutting down safely and being modified, even though doing so could render it unable to fulfill its original goals (Soares et al. 2015).

16. As judicial AI provides increasing support and transitions to fully replace human judges, the role of humans in judicial decision-making might span interruption in order to correct and improve performance by responding to error and bias, to ensure resilience in case of failure or emergency, and to increase system legitimacy, among others (see Crootof et al. forthcoming, pp. 45–61).

REFERENCES

Alarie, Benjamin. 2016. "The Path of the Law: Towards Legal Singularity," *University of Toronto Law Journal*, vol. 66, no. 4, pp. 443–455.

Aleinikoff, T. Alexander. 1987. "Constitutional Law in the Age of Balancing," *Yale Law Journal*, vol. 96, no. 5, pp. 943–1005.

Amodei, Dario, and Jack Clark. 2016. "Faulty Reward Functions in the Wild," OpenAI, December 21. http://openai.com/blog/faulty-reward-functions/.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." (unpublished manuscript, July 25). http://arxiv.org/abs/1606.06565.

"Artificial Intelligence Incident Database." 2022. Partnership on AI. http://incidentdatabase.ai/.

Atkinson, Katie, Trevor Bench-Capon, and Danushka Bollegala. 2020. "Explanation in AI and Law: Past, Present and Future," *Artificial Intelligence*, vol. 289, December, p. 103387.

Bajgar, Ondrej, and Horenovsky, Jan. 2022. "Negative Human Rights as a Basis for Long-term AI Safety and Regulation." (unpublished manuscript, August 31). https://arxiv.org/abs/2208.14788.

Banks, Christopher P., and David M. O'Brien. 2021. *The Judicial Process: Law, Courts, and Judicial Politics* (West Academic Publishing).

Batarseh, Feras A., Laura Freeman, and Chih-Hao Huang. 2021. "A Survey on Artificial Intelligence Assurance," *Journal of Big Data*, vol. 8, no. 1, art. 60.

Baum, Lawrence. 2009. *Judges and Their Audiences: A Perspective on Judicial Behavior* (Princeton University Press).

Becker, Daniel, and Isabela Ferrari. 2020. "Victor, the Brazilian Supreme Court's Artificial Intelligence: A Beauty or a Beast?," in *Regulação 4.0*, 2020th ed., Vol. II (São Paulo: Editora Revista dos Tribunais).

Bednar, Jenna. 2004. "Judicial Predictability and Federal Stability: Strategic Consequences of Institutional Imperfection," *Journal of Theoretical Politics*, vol. 16, no. 4, pp. 423–446.

Beijing Internet Court. 2019. "White Paper on the Application of Internet Technology in Judicial Practice," Anniversary Series Paper, August 17. http://www.chinadaily.com.cn/specials/WhitePaperontheApplicationofInternetTechnologyinJudicialPractice.pdf.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44.

Berman, Emily. 2018. "A Government of Laws and Not of Machines," *Boston University Law Review*, vol. 98, pp. 1277–1355.

Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy," in "Proceedings of the 1st Conference on Fairness, Accountability and Transparency," special issue, *Proceedings of Machine Learning Research*, vol. 81, pp. 149–159.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." (unpublished manuscript, April 20). http://arxiv.org/abs/2004.07213.

CEPEJ (European Commission for the Efficiency of Justice). 2018. "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment," Council of Europe, December 3–4. http://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

Chander, Anupam. 2017. "The Racist Algorithm?," *Michigan Law Review*, vol. 115, no. 6, pp. 1023–1045.

Chen, Daniel L. 2019. "Machine Learning and the Rule of Law," in *Law as Data*, ed. Michael A. Livermore and Daniel N. Rockmore (SFI Press), pp. 433–441.

Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values* (W. W. Norton & Company).

Christiano, Paul. 2016. "Prosaic AI Alignment," AI Alignment, November 18. http://ai-alignment.com/prosaic-ai-control-b959644d79c2.

———. 2018. "Clarifying 'AI Alignment," AI Alignment, April 7. http://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6

Christiansen, Matthew R., and William N. Eskridge, Jr. 2014. "Congressional Overrides of Supreme Court Statutory Interpretation Decisions, 1967–2011," *Texas Law Review*, vol. 92, no. 6, pp. 1317–1541.

Clouser, K. Danner, and Bernard Gert. 1990. "A Critique of Principlism," *Journal of Medicine and Philosophy*, vol. 15, no. 2, pp. 219–236.

Cohen, Michael K., Marcus Hutter, and Michael A. Osborne. 2022. "Advanced Artificial Agents Intervene in the Provision of Reward," *AI Magazine*, vol. 43, no. 3, pp. 282–293.

Coglianese, Cary, and David Lehr. 2019. "Transparency and Algorithmic Governance," *Administrative Law Review*, vol. 71, no. 1, pp. 1–56.

Crootof, Rebecca. 2019. "'Cyborg Justice' and the Risk of Technological-Legal Lock-In," *Columbia Law Review Forum*, vol. 119, no. 7, pp. 233–251.

Deakin, Simon, and Christopher Markou. 2020. "From Rule of Law to Legal Singularity," in *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence*, ed. Simon Deakin and Christopher Markou (Hart Publishing), pp. 1–30.

Deeks, Ashely S. 2019. "The Judicial Demand for Explainable Artificial Intelligence," *Columbia Law Review*, vol. 119, no. 7, pp. 1829–1850.

———. 2020. "Secret Reason-Giving," *Yale Law Journal*, vol. 129, no. 3, pp. 612–689.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." (unpublished manuscript, March 2). http://arxiv.org/abs/1702.08608.

Eckersley, Peter. 2019. "Impossibility and Uncertainty Theorems in AI Value Alignment." (unpublished manuscript, March 5). http://arxiv.org/abs/1901.00064.

Edwards, Lilian, and Michael Veale. 2017. "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For," *Duke Law & Technology Review*, vol. 16, no. 1, pp. 18–84.

Ellis III, T. S. 2008. "Sealing, Judicial Transparency and Judicial Independence," *Villanova Law Review*, vol. 53, no. 5, pp. 939–950.

Epstein, Lee, William M. Landes, and Richard A. Posner. 2013. *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice* (Cambridge, MA: Harvard University Press).

European Commission. 2021. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final, April 21, http://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

Everitt, Tom, Gary Lea, and Marcus Hutter. 2018. "AGI Safety Literature Review," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Survey track*, pp. 5441–5449.

Everitt, Tom, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. "Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective," *Synthese*, vol. 198 (suppl. 27), pp. 6435–6467.

Everitt, Tom, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. 2021. "Agent Incentives: A Causal Perspective," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, pp. 11487–11495.

Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437.

Garoupa, Nuno, and Tom Ginsburg. 2009. "Judicial Audiences and Reputation: Perspectives from Comparative Law," *Columbia Journal of Transnational Law*, vol. 47, no. 3, pp. 451–490.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. "Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts," *Journal of Artificial Intelligence Research*, vol. 62, no. 1, pp. 729–754.

Green, Kesten C., and J. Scott Armstrong. 2007. "Global Warming: Forecasts by Scientists Versus Scientific Forecasts," *Energy & Environment*, vol. 18, no. 7+8, pp. 997–1021.

Green, Leslie. 2016. "Law and the Role of a Judge," in *Legal, Moral, and Metaphysical Truths: The Philosophy of Michael S. Moore*, ed. Kimberly Kessler Ferzan and Stephen J. Morse (Oxford University Press), pp. 323–342.

Greenstein, Stanley. 2021. "Preserving the Rule of Law in the Era of Artificial Intelligence (AI)," *Artificial Intelligence and Law*.

Grey, Thomas C. 1975. "Do We Have an Unwritten Constitution?," *Stanford Law Review*, vol. 27, no. 3, pp. 703–718.

Hacker, Philipp. 2018. "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law," *Common Market Law Review*, vol. 55, no. 4, pp. 1143–1185.

Hadfield-Menell, Dylan, and Gillian Hadfield. 2018. "Incomplete Contracting and AI Alignment." (unpublished manuscript, April 12). http://arxiv.org/abs/1804.04268

Hayward, Keith J., and Maas, Matthijs M. 2021. "Artificial Intelligence and Crime: A Primer for Criminologists," *Crime, Media, Culture*, vol. 17, no. 2, pp. 209–233.

Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. "Aligning AI with Shared Human Values," International Conference on Learning Representations 2021. http://arxiv.org/abs/2008.02275.

Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. "Unsolved Problems in ML Safety." (unpublished manuscript, June 16). http://arxiv.org/abs/2109.13916

Hilton, Benjamin. 2022. Preventing an AI-related Catastrophe. *80,000 Hours*, August 25. http://80000hours.org/problem-profiles/artificial-intelligence/.

HLR (Harvard Law Review). 2017. "State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing," *Harvard Law Review*, vol. 130, pp. 1530–1537.

Hubinger, Evan. 2022. "A Transparency and Interpretability Tech Tree," *Alignment Forum*, June 16. http://www.alignmentforum.org/posts/nbq2bWLcYmSGup9aF/a-transparency-and-interpretability-tech-tree.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. "Risks from Learned Optimization in Advanced Machine Learning Systems." (unpublished manuscript, June 5). http://arxiv.org/abs/1906.01820.

Huq, Aziz Z. 2020a. "A Right to a Human Decision," *Virginia Law Review*, vol. 106, no. 3, pp. 611–688.

———. 2020b. "Constitutional Rights in the Machine-Learning State," *Cornell Law Review*, vol. 105, no. 7, pp. 1875–1953.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, The. 2019. "Law," in *Ethically Aligned Design* (1st edition), IEEE, pp. 211–281.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. 2018. "AI Safety Via Debate." (unpublished manuscript, October 22). http://arxiv.org/abs/1805.00899.

Jauhar, Ameen, Vaidehi Misra, Arghya Sengupta, Partha P. Chakrabarti, Saptarishi Ghosh, and Kripabandhu Ghosh. 2021. “Responsible AI for the Indian Justice System,” Vidhi Centre for Legal Policy; TCG-Crest.

Kanal, Laveen N., and John F. Lemmer, ed. 1985. *Uncertainty in Artificial Intelligence* (North-Holland).

Kaplow, Louis, and Steven Shavell. 2001. “Fairness Versus Welfare,” *Harvard Law Review*, vol. 114, no. 4, pp. 961–1388.

Karnofsky, Holden. 2021a. “AI Timelines: Where the Arguments, and the ‘Experts,’ Stand,” Cold Takes, September 7. https://www.cold-takes.com/where-ai-forecasting-stands-today/.

———. 2021b. “Forecasting Transformative AI: What’s the Burden of Proof?,” Cold Takes, August 17. http://www.cold-takes.com/forecasting-transformative-ai-whats-the-burden-of-proof/.

Kazim, Emre, and Adriano Koshiyama. 2020. “AI Assurance Processes,” (unpublished manuscript, October 20). http://doi.org/10.2139/ssrn.3685087.

Kenton, Zachary, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. “Alignment of Language Agents.” (unpublished manuscript, March 26). http://arxiv.org/abs/2103.14659.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. “Inherent Trade-Offs in the Fair Determination of Risk Scores,” *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, art. No. 43, pp. 43:1–43:23.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions,” *Quarterly Journal of Economics*, vol. 133, no. 1, pp. 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018. “Discrimination in the Age of Algorithms,” *Journal of Legal Analysis*, vol. 10, pp. 113–174.

Kohli, Pushmeet, Krishnamurthy (Dj) Dvijotham, Jonathan Uesato, and Sven Gowal. 2019. “Identifying and Eliminating Bugs in Learned Predictive Models,” *DeepMind Blog*, March 28. http://deepmind.com/blog/article/robust-and-verified-ai.

Krakovna, Victoria. “Specification Gaming Examples in AI—Master List.” (April 2, 2018). http://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

Landau, David E. 2017. “Substitute and Complement Theories of Judicial Review,” *Indiana Law Journal*, vol. 92, no. 4, pp. 1283–1327.

Langosco, Lauro Langosco Di, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. 2022. “Goal Misgeneralization in Deep Reinforcement Learning,” *Proceedings of Machine Learning Research*, vol. 162, pp. 12004–12019.

Lehman, Joel, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, et al. 2020. “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities,” *Artificial Life*, vol. 26, no. 2, pp. 274–306.

Leike, Jan, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. “Scalable Agent Alignment Via Reward Modeling: A Research Direction.” (unpublished manuscript, November 19). http://arxiv.org/abs/1811.07871.

Maas, Matthijs M. 2022. “AI, Governance Displacement, and the (De)Fragmentation of International Law.” (unpublished manuscript, March 1). http://dx.doi.org/10.2139/ssrn.3806624.

Manheim, David, and Scott Garrabrant. 2019. “Categorizing Variants of Goodhart’s Law.” (unpublished manuscript, February 24). http://arxiv.org/abs/1803.04585.

Martínez, Eric, and Christoph Winter. 2022. “Automating the Judiciary: A Global Survey of Legal Academics.” (unpublished manuscript).

Martínez, Eric, and Kevin Tobia. 2022. “What Do Law Professors Believe about Law and the Legal Academy? An Empirical Inquiry.” (unpublished manuscript, August 8). http://dx.doi.org/10.2139/ssrn.4182521.

Muehlhauser, Luke. 2016. What Should We Learn from Past AI Forecasts?. *Open Philanthropy*, May 1, last updated September 2016. http://www.openphilanthropy.org/research/what-should-we-learn-from-past-ai-forecasts/.

Michaels, Andrew C. 2020. “Artificial Intelligence, Legal Change, and Separation of Powers,” *University of Cincinnati Law Review*, vol. 88, no. 4, pp. 1083–1103.

Niiler, Eric. 2019. “Can AI Be a Fair Judge in Court? Estonia Thinks So.” 2019. *Wired*, March 25. http://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/.

Ngo, Richard. 2020. “AGI Safety from First Principles.” (unpublished manuscript, September). http://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view.

———. 2022a. “The Alignment Problem from a Deep Learning Perspective.” (unpublished manuscript, August 30). http://arxiv.org/abs/2209.00626.

———. 2022b. “Outer vs Inner Misalignment: Three Framings,” Alignment Forum, July 6. http://www.alignmentforum.org/posts/poyshiMEhJsAuifKt/outer-vs-inner-misalignment-three-framings-1.

Olah, Chris. 2022. “Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases,” *Transformer Circuits Thread* (June 27). http://www.transformer-circuits.pub/2022/mech-interp-essay/index.html.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. “Zoom In: An Introduction to Circuits,” *Distill*. http://doi.org/10.23915/distill.00024.001.

Ortega, Pedro A, Vishal Maini, and DeepMind safety team. 2018. “Building Safe Artificial Intelligence: Specification, Robustness, and Assurance,” *DeepMind Safety Research*, September 27. http://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt. 2022. “The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models.” (unpublished manuscript, February 14). http://arxiv.org/abs/2201.03544.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press).

———. 2019. “A Rule of Persons, Not Machines: The Limits of Legal Automation,” *George Washington Law Review*, vol. 87, pp. 1–55.

Pasquale, Frank, and Glyn Cashwell. 2018. “Prediction, Persuasion, and the Jurisprudence of Behaviourism,” *University of Toronto Law Journal*, vol. 68, supplement 1, pp. 63–81.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton, NJ: Princeton University Press).

Tetlock, Philip E., and Gardner, Dan. 2015. *Superforecasting: The Art and Science of Prediction* (New York: Broadway Books).

Posner, Richard A. 2003. *Economic Analysis of Law*, 6th ed. (New York: Aspen).

Re, Richard M., and Alicia Solow-Niederman. 2019. “Developing Artificially Intelligent Justice,” *Stanford Technology Law Review*, vol. 22, no. 2, pp. 242–289.

Rudin, Cynthia. 2019. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” *Nature Machine Intelligence*, vol. 1, pp. 206–216.

Rudin, Cynthia, and Joanna Radin. 2019. “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From an Explainable AI Competition,” *Harvard Data Science Review*, vol. 1, no. 2.

Rudner, Tim G. J., and Helen Toner. 2021. “Key Concepts in AI Safety: Specification in Machine Learning,” Center for Security and Emerging Technology, December. http://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf.

Ruf, Boris, and Marcin Detyniecki. 2021. “Towards the Right Kind of Fairness in AI.” (unpublished manuscript, September 30). http://arxiv.org/abs/2102.08453.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking).

Samaha, Adam M. 2008. “Judicial Transparency in an Age of Prediction,” *Villanova Law Review*, vol. 53, no. 5, pp. 829–854.

SCPI (Stanford Center on Poverty and Inequality). 2017. “State of the Union,” special issue, *Pathways: A Magazine on Poverty, Inequality, and Social Policy*.

Selbst, Andrew D., and Solon Barocas. 2018. “The Intuitive Appeal of Explainable Machines,” *Fordham Law Review*, vol. 87, no. 3, pp. 1085–1139.

Sentencing Project, The. 2000. “Reducing Racial Disparity in the Criminal Justice System: A Manual for Practitioners and Policymakers.”

Shetreet, Shimon. 2012. “Fundamental Values of the Justice System,” *European Business Law Review*, vol. 23, no. 1, 61–75.

Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2019. “Goal Misgeneralization: Why Correct Specifications Aren’t Enough For Correct Goals.” (unpublished manuscript, October 4). http://arxiv.org/abs/2210.01790.

Shi, Changqing, Tania Sourdin, and Bin Li. 2021. “The Smart Court—A New Pathway to Justice in China?,” *International Journal for Court Administration*, vol. 12, no. 1, 4.

Skalse, Joar, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. “Defining and Characterizing Reward Hacking.” (unpublished manuscript, September 27). http://arxiv.org/abs/2209.13085.

Soares, Nate. 2018. “The Value Learning Problem,” in *Artificial Intelligence Safety and Security*, ed. Roman V. Yampolskiy (Boca Raton, FL: Chapman and Hall/CRC), pp. 89–97.

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. 2015. “Corrigibility,” *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (January 25–26).

Sourdin, Tania. 2018. “Judge v Robot? Artificial Intelligence and Judicial Decision-Making,” *University of New South Wales Law Journal*, vol. 41, no. 4, pp. 1114–1133.

———. 2021. *Judges, Technology and Artificial Intelligence: The Artificial Judge* (Cheltenham, UK: Edward Elgar Publishing).

Sourdin, Tania, and Archie Zariski. 2013. *The Multi-tasking Judge: Comparative Judicial Dispute Resolution* (Pyrmont, NSW: Thomson Reuters).

Spiller, Pablo T., and Rafael Gely. 2008. “Strategic Judicial Decision-making,” in *The Oxford Handbook of Ethics of AI*, ed. Gregory A. Caldeira, R. Daniel Kelemen, and Keith E. Whittington (Oxford University Press).

Stern, Rachel E., Benjamin L. Liebman, Margaret E. Roberts, and Alice Z. Wang. 2021. “Automating Fairness? Artificial Intelligence in the Chinese Courts,” *Columbia Journal of Transnational Law*, vol. 59, pp. 515–553.

Sunstein, Cass R. 2019. “Algorithms, Correcting Biases,” *Social Research: An International Quarterly*, vol. 86, no. 2, pp. 499–511.

Surden, Harry. 2020. “Ethics of AI in Law: Basic Questions,” in *The Oxford Handbook of Political Science*, ed. Markus D. Dubber, Frank Pasquale, and Sunit Das (Oxford University Press).

Tamanaha, Brian Z. 2004. *On the Rule of Law: History, Politics, Theory* (Cambridge, UK: Cambridge University Press).

Tetlock, Philip E. 2003. “Thinking the Unthinkable: Sacred Values and Taboo Cognitions,” *TRENDS in Cognitive Sciences*, vol. 7, no. 7, pp. 320–324.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation,” *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99.

Waldron, Jeremy. 1989. “Rights in Conflict,” *Ethics*, vol. 99, no. 3, pp. 503–519.

———. 2006. “The Core of the Case Against Judicial Review,” *Yale Law Journal*, vol. 115, no. 6, pp. 1346–1406.

Walsh, Toby. 2018. “Expert and Non-expert Opinion About Technological Unemployment,” *International Journal of Automation and Computing*, vol. 15, no. 5, pp. 637–642.

Wang, Ran. 2020. "Legal Technology in Contemporary USA and China," *Computer Law & Security Review*, vol. 39, 105459.

Weller Adrian. 2019. "Transparency: Motivations and Challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, vol. 11700, ed. Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert Müller (Cham: Springer), pp. 23–40.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200 (Honolulu, HI: ACM).

Wilson, James G. 1995. "Surveying the 'Forms of Doctrine' on the Bright Line Balancing Test Continuum," *Arizona State Law Journal*, vol. 27, no. 3, pp. 773–843.

Winter, Christoph K. 2020. "The Value of Behavioral Economics for EU Judicial Decision-Making," *German Law Journal*, vol. 21, no. 2, pp. 240–264.

———. 2022. "The Challenges of Artificial Judicial Decision-Making for Liberal Democracy," in *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives*, ed. Piotr Bystranowski, Bartosz Janik, and Maciej Próchnicki (Springer Nature).

Winter, Christoph, Jonas Schuett, Eric Martínez, Suzanne Van Arsdale, Renan Araújo, Nick Hollman, Jeff Sebo, Andrew Stawasz, Cullen O'Keefe, and Giuliana Rotola. 2021. "Legal Priorities Research: A Research Agenda," Legal Priorities Project. http://www.legalpriorities.org/research_agenda.pdf.

Xu, Xiaobing, and George Wilson. 2006. "On Conflict of Human Rights," vol. 5, no. 1, pp. 31–57.

Zalnieriute, Monika, Lyria Bennett Moses, and George Williams. 2019. "The Rule of Law and Automation of Government Decision-Making," *Modern Law Review*, vol. 82, no. 3, pp. 425–455.

Završnik, Aleš. 2020. "Criminal Justice, Artificial Intelligence Systems, and Human Rights," *ERA Forum*, vol. 20, no. 4, pp. 567–583.

Zhang, Baobao, Noemi Dreksler, Markus Anderljung, Lauren Kahn, Charlie Giattino, Allan Dafoe, and Michael C. Horowitz. 2022. "Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers." (unpublished manuscript, June 8). http://arxiv.org/abs/2206.04132.

Zhuang, Simon, and Dylan Hadfield-Menell. 2020. "Consequences of Misaligned AI," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 15763–15773.