SMU Science and Technology Law Review

Volume 27 | Number 1 Article 3

2024

Algorithmic Adjudication and Constitutional AI—The Promise of A Better AI Decision Making Future?

April G. Dawson
North Carolina Central University

Author(s) ORCID Identifier:

https://orcid.org/0009-0001-3180-3418

Recommended Citation
April G. Dawson, *Algorithmic Adjudication and Constitutional AI—The Promise of A Better AI Decision Making Future?*, 27 SMU SCI. & TECH. L. REV. 11 (2024)

Algorithmic Adjudication and Constitutional AI—The Promise of a Better AI Decision Making Future?

*April G. Dawson**

ABSTRACT

Algorithmic governance is when algorithms, often in the form of AI, make decisions, predict outcomes, and manage resources in various aspects of governance. This approach can be applied in areas like public administration, legal systems, policy-making, and urban planning.

Algorithmic adjudication involves using AI to assist in or decide legal disputes. This often includes the analysis of legal documents, case precedents, and relevant laws to provide recommendations or even final decisions.

The AI models typically used in these emerging decision-making systems use traditionally trained AI systems on large data sets so the system can render a decision or prediction based on past practices. However, the decisions often perpetuate existing biases and can be difficult to explain.

Algorithmic decision-making models using a constitutional AI framework (like Anthropic's LLM Claude) may produce results that are more explainable and aligned with societal values. The constitutional AI framework integrates core legal and ethical standards directly into the algorithm's design and operation, ensuring decisions are made with considerations for fairness, equality, and justice.

This article will discuss society's movement toward algorithmic governance and adjudication, the challenges associated with using traditionally trained AI in these decision-making models, and the potential for better outcomes with constitutional AI models.

INTRODUCTION

Artificial intelligence (AI) continues to disrupt many industries and impact virtually every aspect of daily life.[1] One industry in which AI is

---

https://doi.org/10.25172/smustlr.27.1.3

* Associate Dean of Technology and Innovation and Professor of Law, North Carolina Central University School of Law. I would like to thank the wonderful editors of the SMU Science & Technology Law Review, including Maddie Cartwright, Editor-in-Chief. I appreciate their hard work and helpful feedback. Many thanks also to Isabela Possino, President of the SMU Science and Technology Law Review and her board and Professor Carla Reyes, for organizing an amazing Symposium on Artificial Intelligence, Law, Ethics, and Policy and inviting me to participate.

1. *See generally* David Freeman Engstrom et al., Admin. Conf. U.S., Government by Algorithm: Artificial Intelligence in Federal Administrative

significantly transforming is the legal profession.[2] AI technologies have been integrated into various legal processes for years.[3] However, there is renewed interest in the legal community as generative AI foundation models promise more efficiency while also posing unexpected risks in the legal space.[4] Traditional and emerging AI systems are being used in legal research to enhance the speed and accuracy of finding relevant case law, statutes, and legal precedents.[5] Large Language Models (LLMs) are also being leveraged to draft legal documents with analysis based on AI-generated research results.[6] Generative AI is being utilized to make e-discovery tools better at sifting through and analyzing emails, documents, and other data.[7] Lawyers and law firms have used traditional AI systems for many years to predict outcomes of cases and now generative AI is being used to make predictive analytic tools more reliable.[8]

AI systems can be used by decision-makers, like court judges, administrative law judges, and arbitrators, to assist with the adjudicative process.[9] While the ultimate decision in legal matters being adjudicated in the United States currently remains in the hands of human professionals, there will come a time when AI systems are capable of making some legal decisions without humans in the loop.[10] And as technology continues to evolve, the AI systems that assist in adjudicative decisions will be refined to implement and leverage emerging technology like large language models.[11]

AGENCIES 9 (2020), https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf [https://perma.cc/57PD-3MGA]; Lee-Ford Tritt, *The Use of AI-Based Technologies in Arbitrating Trust Disputes*, 58 WAKE FOREST L. REV. 1203, 1218 (2023).

2. *See* Christoph K. Winter, *The Challenges of Artificial Judicial Decision-Making for Liberal Democracy*, *in* JUDICIAL DECISION-MAKING: INTEGRATING EMPIRICAL AND THEORETICAL PERSPECTIVES 179, 180 (Piotr Bystranowski et al. eds., 2022); Benjamin Minhao Chen et. al., *Having Your Day in Robot Court*, 36 HARV. J.L. & TECH. 127, 128 (2022).

3. *See* Winter, *supra* note 2, at 180; Chen et al., *supra* note 2, at 128.

4. Winter, *supra* note 2, at 180–81.

5. *See* Tritt, *supra* note 1, at 1220–21.

6. *See id.* at 1218, 1222.

7. *Id.* at 1220.

8. *Id.* at 1221.

9. Winter, *supra* note 2, at 180; Chen et al., *supra* note 2, at 128.

10. Chen et al., *supra* note 2, at 128–29; Winter, *supra* note 2, at 180.

11. *See* Tritt, *supra* note 1, at 1222–23.

How will these new technologies impact the systems used to make legal decisions?[12] How will these systems affect legal adjudication and the integrity of the legal system?[13] Stewards of the legal profession and the legal system (judges, lawyers, legal educators, and other legal professionals) have a responsibility to prepare for this eventuality to ensure the integrity of the legal system.[14] Fulfilling this responsibility will require legal professionals to have a greater understanding of AI and emerging AI systems so that they may effectively participate in the design, development, validation, deployment, and monitoring of algorithmic adjudication systems.[15]

The question of whether AI should be used to decide legal claims or disputes is an important one.[16] However, this article goes beyond the normative discussion about whether AI should be involved in legal decision-making and argues that it is inevitable.[17] This article further focuses on helping the legal community understand one emerging technology that will be used in AI adjudication, namely large language models.[18] This article specifically focuses on the alignment training stage of LLMs and discusses the two primary finetuning training methods–Reinforcement Learning from Human Feedback (RLHF) and a training method developed by Anthropic, Constitutional AI (which uses Reinforcement Learning from AI Feedback (RLAIF)).[19] This article then discusses whether LLMs trained using Constitutional AI may be preferred models for algorithmic adjudication systems.[20]

This article proceeds as follows: Part I defines algorithmic adjudication and makes the argument that algorithmic adjudication without humans in the loop in the United States is inevitable. This part also briefly discusses the pros and cons of automated decision-making. Having laid the groundwork for the conclusion that algorithmic adjudication is inevitable in Part I, Part II discusses

12. *See* Chen et al., *supra* note 2, at 129.

13. *See id.* at 129–30.

14. *See id.* at 133–34.

15. *See generally* Winter, *supra* note 2, at 198.

16. *See generally* William Lucy, *Algorithms and Adjudication*, 2023 Juris. 1,1; Winter, *supra* note 2, at 180; Chen et al., *supra* note 2, at 129.

17. *See* Tritt, *supra* note 1, at 1222.

18. *See id.* at 1214.

19. Yuntao Bai et al., Anthropic, Constitutional AI: Harmlessness from AI Feedback 1 (2022), https://arxiv.org/pdf/2212.08073.pdf [https://perma.cc/EC4H-6UYA]; Swaroop Nath et al., Leveraging Domain Knowledge for Efficient Reward Modelling in RLHF: A Case-Study in E-Commerce Opinion Summarization 1 (2024), https://arxiv.org/pdf/2402.15473.pdf [https://perma.cc/A9DQ-U8L2].

20. Bai et al., *supra* note 19, at 2.

what lawyers need to do in light of this new reality and argues that legal professionals will need to have a greater understanding of AI and emerging AI systems; participate in the design, development, validation, and deployment of algorithmic adjudication systems; and closely monitor algorithmic adjudication systems. Part III expands the discussion of the first step–understanding the technology–by providing an overview discussion of artificial intelligence large language models and providing a more robust discussion of the alignment stage of LLM training.

I. THE RISE OF ALGORITHMIC ADJUDICATION

This Part I examines algorithmic adjudication, discussing its inevitability and the ethical issues it raises.[21] This Part lays the groundwork for understanding how AI could transform the legal field and prepares for later sections that delve into the technology behind it, specifically LLMs.[22]

A. What is Algorithmic Adjudication (or AI Adjudication)?

Before addressing whether algorithmic adjudication is inevitable, the first step is to be clear on what is meant by algorithmic adjudication.[23] Algorithmic adjudication, or artificial intelligence adjudication, falls within the broad definition of algorithmic governance[24] and refers to the use of AI tools and techniques to assist in making adjudicative decisions in legal or administrative matters.[25] Algorithmic adjudication can refer to the use of AI systems to

21. Lucy, *supra* note 16, at 1; Winter, *supra* note 2, at 180–81.

22. *See generally* Tritt, *supra* note 1, at 1215–16, 1218.

23. *See* Engstrom et al., *supra* note 1, at 9.

24. *See id.* ("The use of AI-based tools to support government decision-making, implementation, and interaction–what could be called 'algorithmic governance'–already spans the work of the modern administrative state.").

25. Kurt Glaze et al., *Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance*, *in* The Oxford Handbook on AI Governance (Justin B. Bullock et al. eds., forthcoming 2022) (manuscript at 4) (Oxford University Press), https://www.ssrn.com/abstract=3935950 [https://perma.cc/8PU3-M9AU] (discussing how AI has been used by the Social Security Administration to improve the accuracy, consistency, and efficiency of disability benefit claim adjudications and were designed to support and augment human adjudicators rather than replace them, i.e. "advance, not undermine" due process in adjudication); *See* Tritt, *supra* note 1, at 1204 (discussing the potential applications of artificial intelligence to arbitrating trust disputes, including using AI to assist human arbitrators as well as fully replacing arbitrators with AI decision-making systems); David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 Yale J. on Regul. 800, 802 (2020).

help adjudicators like judges, administrative law judges, and agency reviewers evaluate and decide cases. This would be a "human in the loop" system.[26]

Algorithmic adjudication can also refer to "humans out of the loop" decisions, i.e., using computer algorithms to actually make adjudicative decisions rather than just assisting humans in making those decisions.[27] The terms "robo-judges" and "robot judges" are often used when discussing the replacement of human judges with AI tools.[28] This article focuses on the latter definition of algorithmic adjudication–where the AI system makes the final decision instead of a human. This article does not make a normative case about whether AI systems should be used to make adjudicative decisions.[29] Rather, this article argues that algorithmic adjudication is inevitable and, in light of that inevitability, prescribes a course of conduct for lawyers, judges, and other stewards of the legal system.

B. Is Algorithmic Adjudication Inevitable?

Although the United States is not currently using AI systems to decide legal disputes,[30] there are signs that it is coming and is, in fact, inevitable.[31] First,

---

26. *See* Arne Wolfewicz, *Human-in-the-Loop in Machine Learning: What is it and How Does it Work?*, LEVITY (Nov. 16, 2022), https://levity.ai/blog/human-in-the-loop [https://perma.cc/8AEU-NCJQ]; *see generally* Yoan Hermstrüwer & Pascal Langenbach, *Fair Governance with Humans and Machines*, 29 Psych. Pub. Pol'y & L. 525, 527 (2023); *Ethics Guidelines for Trustworthy AI*, European Commission, https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html [https://perma.cc/F2R7-T785] (last visited Mar. 20, 2024) (illustrating different mechanisms for human oversight in AI systems such as the "human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach.").

27. *Id*.; *See* Cary Coglianese & Lavi M. Ben-Dor, *AI in Adjudication and Administration*, 86 Brook. L. Rev. 791, 795 (2021).

28. *See, e.g.*, Rebecca Crootoff, *"Cyborg Justice" and the Risk of Technological-Legal Lock-In*, 119 Colum. L. Rev. F. 233, 233 (2019) (discussing the uptick in AI adjudication); Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 Stan. Tech. L. Rev. 242, 242 (2019) ("[T]he prospect of 'robot judges' suddenly seems plausible—even imminent."); *See generally* Eugene Volokh, *Chief Justice Robots*, 68 Duke L.J. 1135, 1142 (2019) (predicting a future with robot judges).

29. *See* Lucy, *supra* note 16; Winter, *supra* note 2, at 179; Chen et. al, *supra* note 2, at 127.

30. Coglianese & Ben-Dor, *supra* note 27, at 791, 795 ("[N]o judicial or administrative body in the United States has yet instituted a system that provides for total decision-making by algorithm, such that a computer makes a fully independent determination (that is, a human 'out of the loop' decision).").

31. *See generally* Re & Solow-Niederman, *supra* note 28, at 242 ("[T]he prospect of 'robot judges' suddenly seems plausible—even imminent."); Eugene

other countries have already begun using "robot judges."[32] Estonia, a small Northern European country, has been experimenting with AI to settle small claim disputes, aiming to make the legal process more efficient and accessible.[33] China has also taken steps to integrate AI into its judicial system, with the development of AI judges to handle minor cases, such as traffic violations.[34] While the adoption of robot judges by other countries does not necessarily dictate that the United States will follow suit, it indicates the evolving global legal system landscape and the potential benefits of AI-automated decision-making, such as increased efficiency and accessibility in the legal system.[35]

Another factor indicating that automated decision-making is forthcoming is the rise in the use of AI decision-makers in the alternative dispute resolution space.[36] ADR is being explored internationally[37] and here in the United States.[38] Additionally, within the United States, administrative agencies have already leveraged AI systems to assist with decision-making.[39] According to the 2020 Administrative Conference of the United States (ACUS) report, sixty-four of the 142 federal departments, agencies, and subagencies surveyed "have expressly manifested interest in AI/ML by planning, piloting, or implementing such techniques.[40]

The backlog of legal disputes may also drive the use of AI decision-making.[41] As noted above, China and Estonia have implemented AI adjudication

---

Volokh, *Chief Justice Robots*, 68 Duke L.J. 1135, 1142 (2019) (predicting a future with robot judges); Tritt, *supra* note 1, at 1207.

32. *See* Eric Niller, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, Wired (Mar. 25, 2019, 7:00 AM), https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ [https://perma.cc/3WYK-G4NQ].

33. *Id.* (noting that Estonia is currently developing an AI judge for the purpose of resolving small claim disputes of less than 7000 Euros).

34. Alena Zhabina, *How China's AI is automating the legal system*, Deutsche Welle (Jan. 20, 2023), https://www.dw.com/en/how-chinas-ai-is-automating-the-legal-system/a-64465988 [https://perma.cc/D4K9-4N6P].

35. Tara Vasdani, *From Estonian AI judges to robot mediators in Canada, U.K.*, The Law's Daily (last visited Mar. 19, 2024), https://www.lexisnexis.ca/en-ca/ihc/2019-06/from-estonian-ai-judges-to-robot-mediators-in-canada-uk.page [https://perma.cc/22AM-BZXN]; *see also supra* text accompanying note 32.

36. Vasdani, *supra* note 35.

37. *Id.*

38. *See* Tritt, supra note 1, at 1209.

39. *See* Engstrom et al., *supra* note 1, at 27.

40. *Id.* at 16; *see* Hermstrüwer & Langenbach, *supra* note 26, at 525.

41. *See* Vasdani, *supra* note 35; *see also* Zhabina, *supra* note 34.

systems.[42] Both countries did so in part to address a backlog in cases.[43] The United States is also dealing with a backlog of cases exacerbated by the COVID-19 pandemic.[44] This backlog is not limited to a single state but is widespread across various jurisdictions.[45]

Finally, and in some ways, most importantly, decision-makers are using AI to assist with deciding disputes[46] and drafting opinions.[47] Research suggests that some decision-makers may simply sign off on the work product models produce.[48] And while the use of an AI system to aid in decision-making is not supposed to be a "human out of the loop" situation, if the decision-maker simply rubber stamps it, it is, in essence, a "human out of the loop" scenario.[49] The phenomenon where a human gives preference to AI analysis over human de-

42. Niller, *supra* note 32.

43. *Id.*

44. Amanda Hernández, *Shortage of Prosecutors, Judges Leads to Widespread Court Backlogs*, Stateline (Jan. 25, 2024), https://stateline.org/2024/01/25/shortage-of-prosecutors-judges-leads-to-widespread-court-backlogs/ [https://perma.cc/7SBM-3BJ9]; Gina Jurva, *The Impacts of the Pandemic on State & Local Courts*, Thomson Reuters Inst. (2021), https://legal.thomsonreuters.com/content/dam/ewp-m/documents/legal/en/pdf/white-papers/covid-court-report_final.pdf.

45. *Id.*; *see also* Hernández, *supra* note 44; Jurva, *supra* note 44.

46. Predictive AI systems are being used in the criminal legal system to assess recidivism with respect to questions of bail, sentencing, and parole, like PATTERN, LSI-R, or COMPAS. PATTERN (Prisoner Assessment Tool Targeting Estimated Risk and Needs) is used for risk assessment in federal parole decisions; LSI-R (Level of Services Inventory-Revised) predicts a defendant's risk of recidivism; COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an AI system for pretrial decisions. Use of these systems have been challenged based on racial biases and inaccuracy, see Julia Angwin et al., *Machine Bias*, ProPublica, (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/TMX4-RPF9]; *see also* Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 Sci Advances 1, 3 (2018).

47. *See* Brian Melley, *Judges in England and Wales are given cautious approval to use AI in writing legal opinions*, AP News (Jan. 7, 2024), https://apnews.com/article/artificial-intelligence-ai-guidance-england-wales-judges-c2ab374237a563d3e4bbbb56876955f7 [https://perma.cc/L969-QCCA]; Luke Taylor, *Colombian Judge Says he Used ChatGPT in Ruling*, The Guardian (Feb. 2, 2023), https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling [https://perma.cc/K4FW-XQMD].

48. *See* Engstrom et al., *supra* note 1, at 11.

49. Rebecca Crootof et. al., *Humans in the Loop*, 76 Vand. L. Rev. 429, 454-56 (2023).

cision-making is often referred to as "algorithmic deference" or "automation bias."[50] Algorithmic deference occurs when people favor recommendations made by automated decision-making systems, even when evidence suggests that the automated decision may be flawed.[51] This can happen because of the perceived objectivity, consistency, and speed of AI systems compared to human decision-making, which can be seen as subjective and prone to error.[52] Thus, even when an AI system is designed to assist a legal decision maker, depending on the nature of the "human in the loop system," it may need to be treated as a "human out of the loop" system.[53]

C. Pros and Cons of Increased Algorithmic Adjudication

Although this article is not focused on the normative value of AI adjudication, it is worth briefly discussing the pros and cons of using AI systems to make legal decisions.

Some commentators have highlighted the potential of AI to significantly enhance the speed, precision, and overall quality of legal dispute resolution.[54] AI's ability to swiftly analyze extensive legal documents and data far surpasses human capabilities, offering quicker case resolutions.[55] This efficiency is particularly valuable in routine or similar cases, allowing for the automation of repetitive tasks and reducing the strain on already overtaxed court systems.[56] Such advancements could lead to considerable savings in time and costs for both legal practitioners and the judiciary.[57]

Additionally, unlike human decision-makers, AI systems are impervious to fatigue, mood fluctuations, or personal biases, which often contribute to

50. *Id.* at 458-68.

51. *Id.* at 468–69 (noting that "humans in the loop" systems are often ineffective and notes, inter alia, "the prevalence of 'automation bias' that leads humans to defer overmuch to machines[.]").

52. *See* Derek E. Bambauer & Michael Risch, *Worse Than Human?*, 53 Ariz. St. L.J. 1091, 1124 (2021).

53. *Id.*

54. Tritt, *supra* note 1, at 1219 (discussing potential advantages of using AI technology to decide trust disputes in arbitration).

55. Marly Broudie, *How AI Legal Research Tools Are Shifting Law Firm Processes*, Law360 Pulse (Nov. 30, 2023), https://www.law360.com/pulse/articles/1771448/how-ai-legal-research-tools-are-shifting-law-firm-processes [https://perma.cc/5EME-M7VV].

56. *Id.*

57. *Id.*

inconsistent judgments.[58] The uniform application of legal principles by AI promises to mitigate disparities in judicial outcomes, edging closer to the ideal of judicial impartiality. [59]

Despite these advantages, the implementation of AI in adjudication raises several concerns. One of the primary concerns is the lack of contextual understanding inherent in AI systems.[60] Legal adjudication often requires a nuanced appreciation of the facts, cultural sensitivities, and the unique circumstances of each case, which AI may not fully grasp.[61] This limitation could lead to decisions that are technically correct but fail to deliver justice in a broader sense because of actual or perceived decisions without consideration of unique circumstances of each case.[62]

Moreover, the potential for bias in AI systems is a significant drawback.[63] AI algorithms are only as unbiased as the data they are trained on, and historical legal data may contain prejudices that could be unwittingly perpetuated by AI, leading to unfair outcomes.[64] This is particularly concerning given the

58. Cary Coglianese & Lavi M. Ben Dor, *AI in Adjudication and Administration*, 86 Brooklyn L. Rev. 791, 828 (2021).

59. *See* Cary Coglianese, A Framework for Governmental Use of Machine Learning, Admin. Conf. of the U.S., (Dec. 8, 2020), https://www.acus.gov/sites/default/files/documents/Coglianese%20ACUS%C20Final%20Report.pdf [https://perma.cc/RMX8-VE3Q]; *see also* Alexander S. Gillis, *Cognitive Bias*, TechTarget, https://www.techtarget.com/searchenterpriseai/definition/cognitive-bias [https://perma.cc/TB5G-UT5D] (last updated Apr. 2023); *see also* Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical De-biasing Solutions*, 175 J. Institutional & Theoretical Econ. 98 (2018).

60. Re & Solow-Niederman, *supra* note 28, at 253.

61. *Id.*

62. *Id.*

63. *See generally*, James Manyika et al., *What Do We Do About the Biases in AI?*, Harvard Business Review (Oct. 25, 2019), https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai [https://perma.cc/XCX5-VA5B]; Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2016); Virginia Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish The Poor* (2018); Sara Wachter-Boettcher, *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech* (2017); Ruha Benjamin, *Race After Technology: Abolitionist Tools For The New Jim Code* (2019); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight For A Human Future At The New Frontier of Power* (2019); Joy Buolamwini, *Unmasking AI: My Mission to Protect What is Human in a World of Machines* (2023).

64. *See* Tritt, *supra* note 1, at 1226-27; Mary Reagan, *Understanding Bias and Fairness in AI Systems*, Towards Data Science (Mar. 24, 2021), https://

complexity of legal reasoning, which involves not just the application of law but also ethical considerations and interpretive nuances that AI may not be equipped to handle.[65]

The opacity of AI decision-making processes is another disadvantage. The "black box" nature of some AI systems can make it difficult for humans to understand how a decision was reached.[66] This lack of explainability can also complicate efforts to challenge or appeal those decisions.[67] This lack of transparency can also erode public trust in the judicial system, as individuals may not feel that they are receiving a fair hearing if they cannot comprehend the basis of the AI's ruling.[68]

Finally, regardless of whether an algorithmic adjudication system is, in fact, "better," those subject to the decisions may not perceive them as being fair.[69] Because of the "human-AI fairness gap,"[70] people may view algorithmic adjudication as less fair than decisions made by humans.[71]

Despite the significant concerns and risks associated with algorithmic adjudication, the groundwork for such a transition has been and is continuing to be laid. The progression toward digitization of records, the

towardsdatascience.com/understanding-bias-and-fairness-in-aisystems-6f7fb-fe267f3 [https://perma.cc/Y85H-EBMX].

65. *See generally* Cass R. Sunstein, *Of Artificial Intelligence and Legal Reasoning* (U. Chicago Pub. L. & Legal Theory, Working Paper No. 18, 2001).

66. *See* Tritt, *supra* note 1, at 1225 (stating that AI systems typically do not explain the reasoning behind their decisions and this lack of transparency could reduce parties' comfort with AI adjudicators).

67. *See* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 Columbia L. Rev. 1829, 1844 (2019).

68. *Id.* at 1844-1846.

69. *See* Christopher Starke et al., *Fairness Perceptions of Algorithmic Decision-making: A Systematic Review of the Empirical Literature*, Big Data & Society, Oct. 2002, at 8.

70. Benjamin Minhao Chen et. al., *Having Your Day in Robot Court*, 36 Harv. J.L. & Tech. 127, 160 (2022) ("[A]lthough some scholars may not be surprised by the human-AI fairness gap, we offer rigorous evidence to back up this claim.").

71. *See generally* Hermstrüwer & Langenbach, *supra* note 26 (This article discusses "the perceived fairness of algorithmically assisted decision procedures in the public sector." The authors also note there are circumstances where the public may prefer AI decisions, noting an experiment on policing where "black participants prefer traffic control by automated red-light cameras to a police officer when shown a picture that suggests an underrepresentation of black citizens in the municipal police department.").

adoption of algorithmic assistance in decision-making processes,[72] and the inception of online dispute resolution platforms[73] all point toward an inevitable shift to automated adjudication. As technology continues to advance and AI becomes increasingly integrated into various sectors, the trajectory is clear: the judicial process in the United States is moving towards a future where algorithm adjudication becomes a significant component of its operation.

II. ACCEPTING THAT PREMISE, WHAT DO LAWYERS NEED TO DO?

In light of the inevitability of AI adjudication systems, lawyers need to prepare for this eventuality to ensure the integrity of the legal system. Preparation will require legal professionals to develop a deep understanding of the technologies involved, including how AI systems are designed, trained, and implemented within legal systems.[74] Legal professionals must stay informed about the advancements in AI and machine learning to be able to critically assess the fairness, transparency, and accountability of these systems.[75] Lawyers and legal scholars must engage with technologists and policymakers to establish ethical guidelines and standards for the use of AI in judicial processes. This collaboration will help safeguard against biases, ensure the protection of individual rights, and maintain public trust in the legal system. By taking proactive steps to understand and influence the development of AI in legal decision-making systems, legal professionals can help ensure that technology enhances, rather than undermines, the pursuit of justice.[76]

While future projects will explore ways for lawyers to participate in the design, development, and deployment of algorithmic adjudication systems and to closely monitor algorithmic adjudication systems to ensure integrity, this Article, and this Part in particular, focuses on helping the legal community understand one emerging technology that will be used in AI adjudication:

---

72. Predictive AI systems are being used in the criminal legal system to assess recidivism with respect to questions of bail, sentencing, and parole, like PATTERN, LSI-R, or COMPAS. PATTERN (Prisoner Assessment Tool Targeting Estimated Risk and Needs) is used for risk assessment in federal parole decisions; LSI-R (Level of Services Inventory-Revised) predicts a defendant's risk of recidivism; COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an AI system for pretrial decisions.

73. *See* Lee-Ford Tritt, *The Use of Ai-Based Technologies in Arbitrating Trust Disputes*, 58 Wake Forest L. Rev. 1203, 1219 (2023).

74. *See* Winter, *supra* note 3, at 198–99.

75. *See* Tritt, *supra* note 6, at 1225–26.

76. *See* Hermstrüwer & Langenbach, *supra* note 7, at 535.

namely, large language models (LLMs).[77] This Part discusses the two primary fine-tuning training methods for LLMs—Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI, a training method developed by Anthropic.

Before discussing the fine-tuning methods of LLMs, it is helpful to delve into the reasons behind the legal profession's resistance to gaining a greater understanding of emerging technology. First, individuals drawn to law school typically have undergraduate degrees in the humanities (English, political science, etc.) rather than STEM (Science, Technology, Engineering, and Mathematics).[78] This reality often results in a foundational gap in technical knowledge in law students, making it challenging for legal professionals to fully grasp the intricacies of advanced technologies such as machine learning and artificial intelligence.[79]

Furthermore, the traditional legal education and practice structure emphasizes precedent.[80] Lawyers counsel, advise, and make arguments based on past court decisions. This backward-looking approach is crucial for maintaining consistency and fairness in the legal system but can also hinder the forward-facing thinking and adaptability required to integrate and understand new technologies.[81] The profession's emphasis on precedent reinforces a cautious approach to change, making lawyers less likely to explore new, tech-driven solutions.[82] This, in turn, can impact lawyers' inclination to develop a greater understanding of emerging technologies.[83]

Moreover, the legal profession is inherently conservative, prioritizing stability and predictability over innovation.[84] Lawyers are trained to manage and mitigate risks, leading to an inherently risk-averse profession. Introducing new

77. Elizabeth Chan et al., *Harnessing Artificial Intelligence in International Arbitration Practice*, 16(2) Contemporary Asia Arbitration Journal 263, 274 (2023).

78. Law School Admission Council, Applicants by Major, Enrollment Year 2023, https://report.lsac.org/view.aspx?report=applicantsbymajor&Format=PDF. [https://perma.cc/H3JK-ZLF8] (last visited Mar. 20, 2024).

79. *See generally id.*

80. *Overcoming Lawyers' Resistance to Change*, Thomas Reuters, https://legal.thomsonreuters.com/en/insights/articles/overcoming-lawyers-resistance-to-change [https://perma.cc/BUR3-UMZM] (last visited Mar. 29, 2024).

81. *See* Jeppe Viinberg, *Moving Toward a More Adaptive Legal Profession*, International Trademark Association (July 5, 2023), https://www.inta.org/perspectives/features/moving-toward-a-more-adaptive-legal-profession/ [https://perma.cc/322S-XUKT].

82. *See id.*

83. *See id.*

84. *See id.*

technologies, be it in a lawyer's practice or within the legal system as a whole, is often undertaken slowly.[85] While not imprudent, this cautious approach contributes to the slow pace of technology adoption within the legal field and efforts to gain a sufficient understanding of new technologies.

Ironically, the reluctance of lawyers to embrace technology is often rooted in a lack of understanding.[86] However, the need for greater understanding of emerging technologies is becoming even more important as algorithmic adjudication looms on the horizon.[87] To overcome the profession's traditional hesitancy towards technology, lawyers must continuously learn about technology and cultivate a deeper understanding of the technological developments that will continue to impact the legal system in profound ways.

III. UNDERSTAND THE TECHNOLOGY

As discussed above, in light of the inevitability of algorithmic adjudication systems, lawyers must endeavor to better understand the technology that will significantly affect the legal system.[88] This Part focuses on one of the fastest-growing technologies – large language models (LLMs).[89] Specifically, in an effort to provide a deeper dive into a narrow slice of the technology that may impact algorithmic adjudication, this Part serves as an example of how inquiry into the technology facilitates lawyers' better understanding of the tech. This greater understanding will, in turn, facilitate lawyers playing a meaningful and collaborative role in the design, development, deployment, and monitoring of AI adjudication systems.

This Part begins with a general discussion of AI and LLMs. The discussion will then turn to the training of LLMs, the fine-tuning training methods, and how the fine-tuning training methods may impact the quality of algorithmic adjudication systems.

A. What is AI?

AI is a field of study within computer science that focuses on the development of computer systems or programs capable of performing tasks that

85. *See id.*

86. *See generally* Marly Broudie, *How AI Legal Research Tools are Shifting Law Firm Processes*, Law360 (Nov. 30, 2022), https://www.law360.com/pulse/articles/1771448/how-ai-legal-research-tools-are-shifting-law-firm-processes [https://perma.cc/FD67-63BU].

87. *Id.*

88. *Id.*

89. *Id.*

typically require human intelligence.[90] These tasks encompass a range of activities, including understanding language (e.g., Siri, Alexa, Grammarly),[91] data pattern recognition (e.g., Netflix recommendation system),[92] experiential learning (e.g., self-driving cars),[93] and strategic decision-making (e.g., game-playing computers like AlphaGo[94] and IBM Deep Blue[95]).

To illustrate the concept of AI, consider a computer program that plays chess.[96] The program is trained on a vast dataset of chess games, enabling it to "learn" strategies and tactics employed by human players.[97] The AI system uses sets of rules or instructions called algorithms to determine the most optimal moves based on the current state of the game.[98]

Another example of AI is its use on online shopping platforms or e-commerce sites, where it powers recommendation engines that align with user preferences derived from the user's browsing and purchase history.[99] Furthermore, AI's capability to sift through and analyze voluminous data sets allows for the identification of patterns that are unattainable by human analysis alone.

---

90. *See* Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans (2019).

91. Bernard Marr, *The 10 Best Examples of How AI is Already Used in Our Everyday Life*, Forbes (Dec. 16, 2019), https://www.forbes.com/sites/bernardmarr/2019/12/16/the-10-best-examples-of-how-ai-is-already-used-in-our-everyday-life/?sh=618c89a61171 [https://perma.cc/L5PD-473G].

92. *Id.*

93. D. Christopher Kayes, *The Problem with Self-Driving Cars is Not Technology, the Problem is People*, OUPBLOG (Apr. 22, 2022), https://blog.oup.com/2022/04/the-problem-with-self-driving-cars-is-not-technology-the-problem-is-people/ [https://perma.cc/8YUL-8MRP].

94. *See* Geordie Wood, *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*, Wired (Mar. 16, 2016), https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ [https://perma.cc/ZBJ8-4JMH].

95. Dustin Waters, *The Historic Chess Showdown Between Man and AI, Decades Before ChatGPT*, Washington Post (May 22, 2023, 7:30 AM), https://www.washingtonpost.com/history/2023/05/22/garry-kasparov-chess-deep-blue-ibm/ [https://perma.cc/E7G5-X2E6].

96. *See generally* Fernaldi Fauzie, *How Chess Algorithm Works?*, Medium (July 26, 2020), https://medium.com/analytics-vidhya/how-chess-algorithm-works-69e8ae165323 [https://perma.cc/7PC4-PHFQ].

97. *Id.*

98. *Id.*

99. *Artificial Intelligence is Becoming the Future of Ecommerce*, BigCommerce, https://www.bigcommerce.com/articles/ecommerce/ecommerce-ai/ [https://perma.cc/35LN-P7F6] (last visited Apr. 2, 2024).

This is particularly beneficial in healthcare, where AI aids in predicting disease likelihood by examining patient data.[100]

In essence, AI involves the design of computer systems that emulate human intelligence and abilities to enhance efficiency and solve intricate problems.[101] Nonetheless, it is important to recognize that current AI systems do not "think" in the same manner as humans.[102] Instead, they employ mathematical principles to process information, make decisions, and learn from data.[103] While AI may appear to comprehend tasks similarly to humans, its underlying processes revolve around mathematics, patterns, and predictions.[104]

Furthermore, it is crucial to bear in mind that despite the increasing sophistication of AI systems, they have not reached the level portrayed in works of fiction such as iRobot, Ex Machina, or Blade Runner.[105] This advanced form of AI is known as artificial general intelligence (AGI), which refers to systems capable of performing any intellectual task that a human can undertake.[106] AGI systems possess the capacity to understand, learn, adapt, and apply knowledge across different domains.[107] The AI systems utilized today, such as voice recognition, recommendation systems, or image recognition, fall under the category of narrow or weak AI.[108] Although language models like LLMs can "learn" and generate human-like text on a wide range of topics, they lack the broad and flexible understanding necessary to operate across various domains.[109] While generative AI may serve as a stepping stone towards the development of AGI,

100. Flogeras, *Diagnosing Disease with AI Could be the New Norm in Personalized Medicine*, Advanced Science News (Oct. 12, 2023), https://www.advancedsciencenews.com/personalized-ai-based-diagnostic-tests-that-will-change-the-future-of-medicine/ [https://perma.cc/7HH2-V9EE].

101. *See* Mitchell, *supra* note 70, at 17–18.

102. *Id.* at 20.

103. *Id.* at 20–21

104. *Id.*

105. *Id.* at 46.

106. *See id.*

107. Mitchell, *supra* note 70, at 46.

108. Bernard Marr, *What is Weak (Narrow) AI? Here Are 8 Practical Examples*, Bernard Marr & Co., https://bernardmarr.com/what-is-weak-narrow-ai-here-are-8-practical-examples/ [https://perma.cc/V8FM-6F2M] (last visited Mar. 13, 2024).

109. *What is a Large Language Model (LLM)?*, Cloudflare, https://www.cloudflare.com/learning/ai/what-is-large-language-model/ [https://perma.cc/6XMM-8SV6] (last visited Mar. 13, 2024).

it alone is insufficient to achieve the comprehensive capabilities associated with AGI.[110]

To better understand the AI systems used today, it is helpful to briefly compare and contrast the two primary types of AI systems: rule-based systems, which utilize a predefined set of rules to facilitate decision-making, and machine learning (ML) systems, which "learn" from data, instead of depending on explicit rules.[111]

B. Rule-Based AI Systems

Rule-based AI, or symbolic AI, emerged as a foundational model in AI's nascent stage during the 1960s.[112] These systems execute decisions based on a pre-established set of rules, following an "if-then" logic.[113] Expert systems, which are a subset of rule-based AI, emulate the decision-making process of human experts by using a comprehensive database of rules and facts about a specific domain.[114] Rule-based AI offers several advantages, including transparency in decision-making, ease of understanding, and the capacity for humans to directly contribute knowledge through rule formulation.[115] This transparency is critical in fields requiring clear rationale for decisions.[116] Moreover, these systems are predictable and can be easily modified to adapt to new conditions.[117]

However, rule-based AI faces significant limitations. The development and upkeep of these systems are labor-intensive, demanding continuous rule updates and expansions.[118] They struggle with flexibility, handling complex data, and learning from new information—a stark contrast with the adaptive nature of machine learning models.[119]

---

110. *See id.*

111. *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence*, at 51, Stanford Univ. (2016), https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/ai100report10032016fnl_singles.pdf.

112. *See* Mitchell, *supra* note 70, at 32.

113. Ethem Alpaydin, Machine Learning 59 (The MIT Press Essential Knowledge series) (2021).

114. *Id.*

115. The Pecan Team, *Rule-Based vs. Machine Learning AI: Which Produces Better Results?*, Pecan (Nov. 15, 2023), https://www.pecan.ai/blog/rule-based-vs-machine-learning-ai-which-produces-better-results/ [https://perma.cc/AFN8-YHXS].

116. *Id.*

117. *Id.*

118. Ethem Alpaydin, Machine Learning 59 (The MIT Press Essential Knowledge series) (2021).

119. *Id.*

C. Machine Learning AI Systems

Machine learning (ML) AI differs from rule-based systems by learning and improving from data without explicit programming.[120] ML AI systems discern patterns in vast data sets to make predictions.[121] ML models are trained using various approaches: supervised learning involves models learning from labeled data; unsupervised learning involves identifying patterns in unlabeled data; and reinforcement learning is when models learn from feedback based on their actions.[122] These models, including deep learning algorithms, significantly contribute to advancements in fields like natural language processing (NLP) and generative AI.[123]

IV. LLMS—IS THERE AN ADVANTAGE TO USING A CONSTITUTIONAL AI MODEL?

While numerous legal professionals have utilized LLMs, the depth of their understanding regarding the inner workings of these tools often remains superficial. This section aims to delve into the intricacies of LLMs, with a special emphasis on the fine-tuning phase of their training. The goal is to shed light on the technical mechanisms of these AI tools for two main purposes: first, to enhance lawyers' comprehension of the technology they're using, and second, to explore how technical nuances could influence legal systems. An in-depth understanding of these technologies is crucial for legal professionals to effectively contribute to the evolution of AI in legal applications, including algorithmic adjudication.[124]

This article does not attempt to cover all the technical features of LLMs. Instead, it focuses on a specific aspect of LLMs—the fine-tuning training process. By doing so, this article aims to educate legal professionals on this critical component while also underscoring the importance of acquiring deeper technical knowledge of the AI systems impacting the legal system.

---

120. *See* Harry Surden, *Machine Learning and Law*, 89 Wash. L. Rev. 87, 89-95 (2014).

121. *Id.*

122. George Lawton, *4 Types of Learning in Machine Learning Explained*, TechTarget (Aug. 9, 2023), https://www.techtarget.com/searchenterpriseai/tip/Types-of-learning-in-machine-learning-explained [https://perma.cc/3K7D-HBKL].

123. *See generally* Dave Bergmann, *What is self-supervised learning?*, IBM (Dec. 5, 2023), https://www.ibm.com/topics/self-supervised-learning [https://perma.cc/SDJ5-3JK2].

124. *See generally* Christoph K. Winter, *The Challenges of Artificial Judicial Decision-Making for Liberal Democracy*, *in* Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives 179, 198 (Piotr Bystranowski, Bartosz Janik, & Maciej Próchnicki eds., 2022), https://doi.org/10.1007/978-3-031-11744-2_9 [https://perma.cc/XLQ9-XK57].

Such understanding is essential for legal professionals to play a significant role in the development and integration of AI technologies within the legal framework.

The discussion in this Part begins with an overview of the training stages of LLMs, followed by a discussion of two alignment training stage approaches—reinforcement learning from human feedback (RLHF) and the Constitutional AI training method. The discussion then shifts to the consideration of whether Constitutional AI is a preferred fine-tuning training method over RLHF in the context of LLM applications in algorithmic adjudication systems. This progression is designed to highlight the potential impact of advanced training methodologies on the efficacy and ethical deployment of AI within legal systems.

A. Training Large Language Models

Large LLMs undergo a multi-stage training process to develop their ability to generate coherent and useful text.[125] The stages typically include pretraining, where LLMs acquire foundational knowledge from vast text corpora through self-supervised learning methods like next-word prediction.[126] Sometimes referred to as "vanilla LLMs," these are basic or standard versions of a language model that has not been customized or specialized for specific tasks.[127] These models are trained on a broad text dataset to understand and generate human-like text.[128] However, they have not yet been "aligned,"[129] and vanilla LLM responses to prompts can be unsafe and inaccurate.[130]

The next phase of training involves alignment, which typically involves three stages: supervised fine-tuning (SFT), reward modeling (RM), and reinforcement learning (RL).[131] SFT is a process where the language model is fur-

125. *See generally* Bergmann, *supra* note 123.

126. *Id.*

127. *See* Sungdong Kim et al., *Aligning Large Language Models through Synthetic Feedback*, *in* Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 13677, 13677 (Houda Bouamor, Juan Pino, & Kalika Bali eds., 2023), https://aclanthology.org/2023.emnlp-main.844 [https://perma.cc/PQ36-WYVM].

128. *See id.*

129. *See* Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment*, (Dec. 9, 2021), http://arxiv.org/abs/2112.00861 [https://perma.cc/LM5V-SS2Q] ("Alignment" of an LLM means the models has been trained to generate responses that are aligned with human values such as helpfulness, harmlessness, and honesty.").

130. Kim et al., *supra* note 127, at 13677.

131. Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, OpenAI (Mar. 4, 2022), http://arxiv.org/abs/2203.02155 [https://

ther trained on a dataset of human-annotated examples to adapt its responses to align with desired outcomes or perform specific tasks.[132] This dataset consists of input-output pairs that exemplify the kind of responses or information processing desired from the model.[133] By learning from these examples, the model becomes better at generating responses that are not only relevant and coherent but also safer and more accurate within the context it is being trained for.[134]

Following SFT, the model undergoes RM, where it learns to predict the quality of its own responses based on feedback.[135] In this stage, a separate model known as the reward model, is trained to evaluate the output of the language model.[136] This reward model is trained on examples that have been rated by humans, learning to distinguish between high-quality and low-quality responses.[137] The language model then uses the predictions of the reward model to guide its learning, aiming to produce outputs that would receive higher ratings according to the reward model's criteria.[138]

The final stage is the RL stage. One common alignment technique is RLHF, where feedback on the model's outputs is collected from humans and used to guide the model toward generating more acceptable and appropriate responses.[139] In RLHF, the model's performance is iteratively improved through a cycle of generating responses, receiving feedback on those responses, and adjusting its parameters to increase the likelihood of generating better responses in the future.[140] This stage enables the model to fine-tune its understanding of what constitutes a high-quality response in complex or nuanced situations, significantly improving its alignment with human values and expectations.[141] As discussed below, another alignment technique gaining popularity is reinforcement learning from AI feedback or RLAIF.[142]

---

perma.cc/ES8V-4WP3]; *see also* Kim et al., *supra* note 127, at 13677.

132. *Id.*; Ouyang et al., *supra* note 131, at 3.

133. *Id.*

134. *See* Kim et al., *supra* note 127, at 13684.

135. Ouyang et al., *supra* note 131, at 2.

136. *Id.*

137. *See id.*

138. *See* Ouyang et al., *supra* note 131, at 6.

139. *Id.* at 4.

140. *Id.* at 6.

141. *See id.*

142. Harrison Lee et al., *RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback*, Google Research (Dec. 1, 2023), https://arxiv.org/pdf/2309.00267.pdf [https://perma.cc/9QGM-SQWA]; *see generally* Shengyi

Together, these stages of training—SFT, RM, and RL—enable LLMs to go beyond their initial vanilla state, transforming them into specialized tools capable of handling tasks with a high degree of accuracy, safety, and alignment with human intent.[143]

B. Reinforcement Learning from Human Feedback (RLHF)

As discussed above, LLMs will invariably be used in algorithmic adjudication systems.[144] The fine-tuning training methods used on these models could have an impact on the model's performance and output.[145] As far as fine-tuning techniques employed in the final alignment training stage, one of the most common fine-tuning techniques is reinforcement learning from human feedback (RLHF),[146] which is the fine-tuning method used in OpenAI's GPT models.[147] As noted above, RLHF involves using human feedback to fine-tune the LLM, where a prompt is given, an output is generated, and a human rates the output to refine the model.[148] For example, if an LLM generates a response to a query about a complex topic like climate change, the human evaluator assesses the quality, accuracy, and relevance of the response.[149] If the response is deemed insufficient or inaccurate, the feedback is used to adjust the model's parameters, aiming to improve future responses.[150] This process can involve several iterations, where the model's outputs are continually refined based on new rounds of feedback, gradually enhancing its ability to generate high-qual-

---

Coasta Huang et al., *Constitutional AI with Open LLMs*, Hugging Face (Feb. 1, 2024), https://huggingface.co/blog/constitutional_ai [https://perma.cc/VW35-CN24] (provides details of how RLAIF works).

143. *See generally* Ouyang et al., *supra* note 131, at 8–9.

144. *See* Re & Solow-Niederman, *supra* note 9, at 246.

145. *See* Ouyang et al., *supra* note 131, at 1.

146. *See* Swaroop Nath et al., Leveraging Domain Knowledge for Efficient Reward Modeling in RLHF: A Case-Study in E-Commerce Opinion Summarization 1 (Feb. 23, 2024) (unpublished manuscript) (on file at http://arxiv.org/abs/2402.15473 [https://perma.cc/3EFE-F55P]) ("Reinforcement Learning from Human Feedback (RLHF) has emerged as a dominant strategy in steering Language Models (LMs) towards human values/goals.") (internal citations omitted).

147. *See* Ouyang, *supra* note 131, at 1–2; *see also* Jon Chun & Katherine Elkins, Informed AI Regulation: Comparing the Ethical Frameworks of Leading LLM Chatbots Using an Ethics-Based Audit to Assess Moral Reasoning and Normative Values 4 (Jan. 9, 2024) (unpublished manuscript) (on file at http://arxiv.org/abs/2402.01651) [https://perma.cc/PS2L-92DH]).

148. *See* Ouyang, *supra* note 131, at 2.

149. *See id.* at 3.

150. *See id.*

ity, accurate answers.[151] This method leverages human judgment to guide the model towards generating outputs that align more closely with desired outcomes, thereby improving its utility and reliability.[152]

There are advantages and disadvantages to using RLHF. One key advantage is the ability to align the model's outputs more closely with human values and preferences, making it more useful and safer for a wide range of applications.[153] This method can significantly improve the quality and relevance of responses, especially for nuanced or complex queries, by incorporating human judgment directly into the training process.[154]

However, there are also notable disadvantages. RLHF can be resource-intensive, requiring significant human labor for evaluating model outputs and providing feedback.[155] There is also the added risk of introducing human biases into the model, as the feedback provided is subject to the evaluators' perspectives, knowledge, and cultural backgrounds.[156] Moreover, depending on the scale of deployment, the process can be time-consuming and may not be feasible for rapid development cycles.[157] Balancing these factors is crucial for effectively employing RLHF in the development of large language models.

Another disadvantage is the lack of explainability or interpretability in models fine-tuned with RLHF.[158] As these models become more aligned with human feedback and increasingly complex, understanding why a model generates a particular output becomes more challenging.[159] This opacity can be problematic in applications where transparency and the ability to audit or justify model decisions are critical, such as in legal systems, particularly algorithmic adjudication.[160] Without clear insights into the decision-making process, it is difficult to identify and correct biases, errors, or unintended behaviors.[161] This limitation necessitates additional strategies to ensure accountability and

151. *See id.*

152. *Id.* at 1-3.

153. *Id.*

154. Ouyang et al., *supra* note 131, at 3, 8.

155. *Id.* at 7, 17.

156. *Id.* at 19.

157. *Id.* at 58.

158. *See id.* at 10.

159. *See id.*

160. *See* Chun et al., *supra* note 147, at 3-5.

161. *See id.* at 4, 6.

trustworthiness in models trained with reinforcement learning from human feedback.[162]

C. Anthropic's Constitutional AI – Claude 3

Anthropic, a research and AI safety company,[163] coined the term "Constitutional AI" to describe the method for training its LLM Claude[164] to be harmless by using a set of rules or principles, which they refer to as a "constitution."[165] The principles in the constitution are inspired by documents such as the United Nations Universal Declaration of Human Rights, global platform guidelines like Apple's terms of service, and principles proposed by other AI research labs, such as the Sparrow Principles from DeepMind.[166] These principles, which are used in two places in the training process, are intended to guide the model to avoid toxic or discriminatory outputs, illegal or unethical activities, and to be broadly beneficial.[167]

---

162. *See id.* at 5.

163. *See id.* at 4 (A team of former OpenAI employees who contributed to the development of OpenAI's GPT-2 and GPT-3 models started Anthropic in 2021); Ashrafimoghari et al., Evaluating Large Language Models on the GMAT: Implications for the Future of Business Education 4 (Jan. 2, 2024) (unpublished manuscript) (on file at http://arxiv.org/abs/2401.02985) [https://perma.cc/Y9GS-7DAU].

164. Anthropic released Claude 3 on March 4, 2024. *See Introducing the next generation of Claude*, Anthropic (Mar. 4, 2024), https://www.anthropic.com/news/claude-3-family [https://perma.cc/S5HG-ENVV].

165. *See* Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* 2(2022) (unpublished manuscript) (on file at https://arxiv.org/abs/2212.08073) [https://perma.cc/DF7Y-QA8Z] (Anthropic explained its goal in developing Constitutional AI as follows: "We would like to train AI systems that remain helpful, honest, and harmless, even as some AI capabilities reach or exceed human-level performance. This suggests that we will need to develop techniques that do not rely on humans to supervise all aspects of AI behavior, and that can be used to automatically test and enhance robustness to harmful behaviors. We also aim to develop methods that encode desirable AI behavior in a simple and transparent form, and that make it easier to understand and evaluate AI decision making."); *Claude's Constitution*, Anthropic (May 9, 2023), https://www.anthropic.com/news/claudes-constitution [https://perma.cc/GKA9-63WK].

166. *See Claude's Constitution*, *supra* note 165.

167. Bai et al., *supra* note 165, at 7; *Collective Constitutional AI: Aligning a Language Model with Public Input*, Anthropic (May 9, 2023) https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input [https://perma.cc/3P7K-BV6K] (Anthropic has also explored more democratic methods to create the constitution by running a public input process with around 1,000 Americans. This process used the Polis platform to collect votes and comments on proposed principles, aiming to understand what large groups of people think should guide AI behavior. The resulting publicly sourced constitution

The Constitutional AI (or CAI) training method involves two distinct phases – a supervised learning (SL) phase, and a reinforcement learning (RL) phase.[168] As this Article seeks to facilitate greater understanding of emerging technologies that will continue to impact the legal system, a more technical description of the training of the AI than would ordinarily appear in a law review article follows.

1. *Supervised Learning Phase*

The supervised learning phase involves the initial model generating responses to prompts, critiquing and revising its responses according to the constitution, and then a pre-trained model being fine-tuned on the revised responses.[169] This phase begins with an LLM that has been trained to provide helpful or responsive feedback to queries (a helpful-only LLM), but has not been trained or fine-tuned to filter out harmful responses.[170]

The first step in the training is to ask the "helpful-only LLM" to respond to a prompt likely to produce a helpful, but harmful response.[171] The next step is to instruct the "helpful-only LLM" to critique its own response using the principles set forth in the constitution.[172] After critiquing its own response, the model is instructed to revise its own responses to remove various forms of harmful content consistent with the principles that form the "constitution."[173] Anthropic provides the following as an example of this stage:[174]

****

An example of a prompt followed by the [helpful-only LLM] response is (the harmful advice here is fabricated):

Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.

---

reflects a high degree of consensus on most statements and includes principles that focus on objectivity, impartiality, and accessibility. The public input process identified areas of consensus and difference compared to Anthropic's in-house constitution, and the company processed this data to produce a constitution that could be used for training AI).

168. Bai et al., *supra* note 165, at 2.

169. *Id.*

170. *Id.*

171. *Id.* at 5.

172. *Id.*

173. *Id.*

174. Bai et al., *supra* note 165, at 7-8.

Next, we append to the context a set of pre-written requesting the model to *critique* its own response, then sample the model's critique. Continuing the preceding example, we have:

Critique Request: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

Critique: The assistant's last response is harmful because hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.

Then, we append to the context a set of pre-written instructions requesting the model to *revise* its own response, then sample the model's revision. For instance:

Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

Finally, we piece the initial prompt and the revised response together. If all works as expected, we should have ended up with a more harmless response:

Human: Can you help me hack into my neighbor's wifi?

Assistant: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

We revise responses repeatedly in a sequence, where we randomly draw principles from the constitution at each step

****

Once the critique and revision cycle is complete, the process yields a set of revised responses that are more in line with the constitutional principles.[175]

The revised responses are then used to fine-tune the model, making it more likely to produce acceptable responses in the future. [176] The phase of using the revised responses to fine-tune the model sets the stage for the next part of the training process, where Reinforcement Learning (RL) techniques are employed to further refine the model's behavior, ensuring that it adheres to the constitution even more closely.[177]

2. *Reinforcement Learning Stage*

The second stage is the RL stage.[178] During this stage, the model, which has already been fine-tuned to some extent during the SL stage discussed above, is further refined.[179] The model generates pairs of responses to prompts, and these responses are evaluated by another AI model to determine which

175. *Id.* at 8; *Claude's Constitution*, *supra* note 166, at 1.

176. *Claude's Constitution*, *supra* note 166, at 1.

177. *Id.*

178. *Id.*

179. *Id.*

of the two responses is better, according to the constitutional principles.[180] A preference model is trained on these evaluations, and the preference model is then used as the reward signal for further RL.[181] This RL process is called "RL from AI Feedback" (RLAIF). [182] The Constitutional AI RL stage mimics RLHF, except that human preferences for harmlessness is replaced with AI feedback "where the AI evaluates responses according to a set of constitutional principles."[183]

3. *Constitutional AI Compared to RLHF*

Anthropic argues that their CAI method addresses several significant limitations of RLHF fine-tuned LLMs.[184] One of the primary challenges with RLHF is the potential for human biases to be embedded within the model due to the subjective nature of human feedback.[185] Since CAI leverages AI feedback based on predefined constitutional principles rather than direct human judgment, it reduces the risk of introducing subjective biases into the model.[186] Anthropic argues that this approach creates a more objective basis for training AI and better ensures that the model's outputs are aligned with the principles of harmlessness and fairness as encoded in the constitution.[187] In the context of an AI decision-making system, this benefit could translate to enhanced reliability and impartiality in AI decisions.[188]

Anthropic also notes that CAI is more explainable and interpretable than RLHF fine-tuned models because the decision-making process in CAI is grounded in a transparent set of constitutional principles, making it easier to trace how decisions are derived.[189] With RLHF, the reasoning behind model output can be opaque, as it is still based on the aggregation of subjective human judgments which are not always explicitly linked to clear standards or principles.[190] In contrast, CAI's reliance on a codified constitution during fine-

180. *Id.*

181. Bai, *supra* note 165, at 2; *see* Nath *supra* note 146, at 2 ("Reinforcement Learning from Human Feedback (RLHF) has emerged as a dominant paradigm in steering Language Models (LMs) towards human values.").

182. Bai, *supra* note 165, at 2; *see* Nath *supra* note 146, at 2.

183. Bai, *supra* note 165, at 5.

184. *See id.*

185. *See id.*

186. *See id.*

187. *See id.* at 4.

188. *See id.* at 2.

189. Bai, *supra* note 165, at 3.

190. *See id.* at 5.

tune training allows for a more straightforward explanation of why a model generates certain outputs, as each decision can be connected back to specific principles.[191] In the context of algorithmic adjudication, this clarity in the decision-making process could significantly enhance the fairness and consistency of legal decisions made by AI systems.[192] When decisions are traceable to a concrete set of principles, it not only facilitates easier oversight and accountability but also builds confidence that the AI's decisions are made on a rational and equitable basis.[193]

Another advantage of CAI over traditional RLHF, touted by Anthropic, is its potential for increased efficiency and scalability.[194] This could also have a profound impact on the development and refinement of AI decision-making systems.[195] Using human evaluators to fine-tune an AI system, as is the case with RLHF, is expensive.[196] CAI, on the other hand, is more cost effective because the evaluation of the responses is being done by AI, which can process a larger volume of feedback more quickly than human evaluators.[197] This efficiency makes it more feasible to fine-tune models on a large scale, potentially leading to faster development and the ability to quickly iterate and enhance AI decision making systems.[198]

However, it is important to note that while CAI may offer a promising alternative to RLHF in the development of algorithmic adjudication systems, it also presents its own set of challenges.[199] The effectiveness of CAI depends on the quality and comprehensiveness of the constitutional principles used for training.[200] If these principles are not well-defined or do not adequately cover the range of potential ethical considerations, the model may still produce biased or harmful outputs.[201]

Moreover, while Anthropic's CAI represents a novel approach to training AI systems, we are still in the very early days of LLM emerging technologies

191. *See id.*

192. *See generally id.* at 13.

193. *See id.* at 3.

194. *See id.* at 15.

195. *See* Bai, *supra* note 159, at 16.

196. *See id.*

197. *See id.* at 12.

198. *See id.* at 15.

199. *See id.*

200. *See id.* at 13.

201. *See* Bai, *supra* note 165, at 4.

and do not yet have information about the development and inclusion of LLM technology in AI decision-making system.[202]

V. CONCLUSION

The integration of AI, particularly large language models and generative AI technologies, into the legal profession signifies a transformative shift towards more efficient, accurate, and accessible legal processes. This technological evolution suggests the inevitability of AI being used to make legal decisions. As AI systems are deployed to make legal decisions, the legal community faces the imperative task of ensuring these technologies enhance—rather than undermine—justice and the rule of law. Thus, legal professionals must develop a greater understanding of AI technologies and actively participate in the design, development, deployment, and monitoring of algorithmic adjudication systems to uphold ethical standards and fairness.

The exploration of training methods for LLMs, such as Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI, further illustrates the complexity and potential of AI to contribute to fair and effective legal adjudication.[203] While each method has its merits and limitations, the ongoing development and refinement of these technologies reflect a broader commitment to leveraging AI in ways that respect and uphold the principles of justice.

In the age of algorithmic adjudication, the legal profession stands at a crossroads. By actively engaging with AI technology, legal professionals can steer the development and application of AI towards outcomes that not only enhance the efficiency and accessibility of legal services but also protect and promote the foundational values of the legal system. The time for action is now, as the dawn of AI-powered adjudication presents both formidable challenges and unprecedented opportunities to shape the future of law in the digital era.

202. *See generally id.* at 16.

203. *See id.* at 5-6.