Rules, Cases, and Reasoning: Positivist Legal Theory as a Framework for Pluralistic AI Alignment

Nicholas A. Caputo[*]
Oxford Martin AI Governance Initiative
Oxford, UK OX1 3BD
`ncaputo@jd24.law.harvard.edu`

Abstract

Legal theory can address two related key problems of alignment: pluralism and specification. Alignment researchers must determine how to specify what is concretely meant by vague principles like helpfulness and fairness and they must ensure that their techniques do not exclude alternative perspectives on life and values. The law faces these same problems. Leading legal theories suggest the law solves these problems through the interaction of rules and cases, where general rules promulgated by a democratic authority are given specific content through their application over time. Concrete applications allow for convergence on practical meaning while preserving space for disagreement on values. These approaches suggest improvements to existing democratic alignment processes that use AI to create cases that give content to rules, promoting more pluralist alignment.

1 Introduction

The law must confront questions of how to reconcile competing values and views of what society is and should be by nature of its very form. It is a set of general rules, made by bodies recognized as having lawmaking authority, that bind all members of a society in the specificity of their daily conduct despite fundamental disagreements on how to live and what matters [27]. These general rules are applied to disputes in cases that, in a system of precedent, bind others in future disputes on similar points of law [24, 12, 46]. How is it possible to preserve a pluralist society that respects the free choice of individuals to make their own decisions given the necessary extent to which general rules reduce individual freedom of choice? And how can general rules predictably be given concrete meaning given the ambiguity of language and changes in circumstances?

Alignment, particularly the kind of finetuning alignment that underlies leading large language models, [2, 8, 33, 34, 41, 43, 37, 47] faces these same problems. Any principle or rule put into a model shapes its outputs in a way that others, including users, might find incompatible with their values, and researchers are still not certain how models interpret or will apply the rules put into them [21, 28, 53, 32, 22]. As Kundu et al. [38] note, simply giving a rule to a model means that the model will interpret the rule based on its underlying black-boxed sense of the words in which the rule is expressed rather than according to what the researchers wanted. Furthermore, there may be significant tradeoffs between alignment to user preferences and to broader social values [35, 10]. Leading approaches are either the product of small groups in AI labs or, where democratic inputs have been sought, of majorities of those few selected to provide such input [11, 3, 23, 42]. Neither approach promotes pluralism. Pluralistic models must instead be able to reason contextually and from different perspectives, drawing on the body of alignment rules and on normal linguistic practice

---

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to ensure that there is consistency and predictability in their outputs but also taking the perspective of minorities with different views from those predominant in training data or in labs [4, 36, 49, 48].

Legal theory provides ideas that might aid in solving these problems. The law specifies its general rules through the application of those rules in concrete cases. In each factual dispute that forms a case, an interpretation of the rule is tested and becomes precedent [5]. Over time, these interpretations accumulate and map out not just the core meaning of a rule but also how to think about its edge cases [27]. Reasoning about concrete facts rather than high level justifications allows pluralism to flourish because members of society can coordinate on concrete ways to live while respecting each other's values [50, 51]. Alignment researchers have begun to draw on legal concepts to align their systems [3, 1, 23, 17, 16, 7, 25, 30, 9]. Incorporating the interplay between rules and cases at the heart of the law can deepen this conversation, improving alignment while promoting democratic pluralism.

2   Positivist Theories for Pluralistic AI Alignment

The law has long confronted the same problems of pluralism and specification that alignment is now facing, and legal scholars have developed a variety of theories for how it should respond to them [27, 29, 6, 15]. Modern American law makes rules at various levels and takes into its scope much of life and the varieties of human interaction, abstracting norms and patterns of behavior [40]. But each law matters in how it is applied in concrete cases involving the rights and obligations of people, and in fact has no real meaning until such applications–advisory opinions are in fact banned in American federal courts to avoid groundless rulemaking [31]. The question then is how to ensure that general rules bind into the future in a way that protects rights and is predictable such that people can live their lives knowing how the law affects them. Two legal theories explored below suggest that cases, applications of a law over time, can fill in the meaning of general law in a way that allows for specification while preserving space for pluralist disagreement over values.

2.1   Specification Through Rules and Cases

Specifying the meaning of legal rules in a predictable and reasonable way is essential to the rule of law and to avoiding arbitrariness and abuses of power [6, 29]. But natural language is ambiguous and contextual. Sometimes bees are legally fish [44]. Professor H.L.A. Hart, likely still the leading scholar of jurisprudence [39], sought to provide an account of adjudication that showed how the meaning of rules is determined in the face of what Hart called this "open texture" of language [27].

The problem that Hart saw was that ambiguity in law persisted despite attempts to add definitions and explanations. Hart's famous example [26] is a sample rule, perhaps made by a town council, that "forbids you to take a vehicle into the public park." At first blush, this rule seems reasonable and easy to understand: it disallows cars and motorcycles from entering the park. But complications arise for cases like bicycles, roller skates, and baby strollers that seem to clearly fit the definition of vehicle but also seem unreasonable to exclude from the park [45]. And what about a military truck intended to be mounted as a memorial [20]? An ambulance coming to the aid of a heart attack victim? Further clarification of the rule is necessary to resolve these cases, and the lengthy law codes risk expanding infinitely more with definitions and exceptions, each with their own definitions and exceptions. This specification problem is of course much worse as soon as the rule at issue moves away from concrete things like vehicles and toward abstractions like justice, helpfulness, bias, or pluralism.

Hart's solution [27] is to say that in language there is a settled core of meaning and a penumbra, at which the edge cases are disputed. When an interpreter of a legal rule is at the core, they should just apply the words at issue as the settled core dictates: no cars in the park. When they are at the fringes, faced with a truly indeterminate case, they must exercise discretion to decide how to interpret the language at issue in light of the social context, underlying goals, or other external features of the situation that might provide guidance. Perhaps there's evidence that the council wanted to avoid accidents from high speeds: then cars are out and bikes might be too, but tricycles are in [27, 15].

But importantly, Hart's system has precedents, past decisions and interpretations that indicate how a new dispute on the same issue should be resolved, and these precedents allow the bootstrapping of the theory of open texture into a relatively complete system of specification that retains a degree of flexibility and contextuality. A case is decided and a precedent interpreting some phrase made. Another dispute about the phrase arises and another decision is made, this time with reference to the

first one. Over time, a kind of constellation of precedent arises around the phrase at issue that allows for definition of the permissible interpretations of the language at issue by means of triangulation among the precedents. The map of meaning is filled out and new cases can relatively easily be fit into the framework of existing law. The original rule at issue gets content through the process of application and people can predict that later interpretations will resemble earlier ones. Thus, specification of meaning is achieved through the interplay of general rules and concrete applications and a law is at once comprehensible and predictable and also flexible enough that it can be applied appropriately in truly novel cases when the need arises [27].

2.2 Analogy and Incompletely Theorized Agreements

Professor Cass Sunstein took the argument a step further, claiming that in fact the law has little need of general principles but rather consists only of analogical reasoning among concrete cases [50], and that this rejection of general principles gives it a special power to protect pluralism [51]. For Sunstein, the engine of the law is analogical reasoning:[2] "fixed points" of precedent are established through judicial decisions and then new cases are decided through incremental analogy to these fixed points [50, 52].[3] Importantly, the justification for the decisions is less relevant than the decisions themselves. Such a low-level, specific approach has important benefits for pluralism. Because the law, at least in adjudication, focuses on concrete cases, there is no need for debates over high level concepts of the good or justice about which people disagree. Instead, two people might agree that a given decision in the case at issue is the right one even if their philosophical justifications for that decision are incompatible [51]. For Sunstein, such a mode of reasoning is necessary for a pluralistic society because it allows for cohabitation among possibly incommensurable values. Unlike Hart, who believed that rules are the core of the law and cases only subsidiary ways of informing their meaning, Sunstein argues that it is the cases that actually function as the core, and that the announcement of general principles risks stamping out differing views.

Alignment on particulars without consensus on principles creates what Sunstein [51] calls "incompletely theorized agreements," which allow for the operation of law in society subject to democratic control and protective of plural perspectives. Incompletely theorized agreements are practical and concrete and provide the substance of the "fixed points," discussed above, from which analogical reasoning to bring new cases within the law can occur. While Sunstein acknowledges that certain higher-level agreements are possible and indeed necessary both on some substantive questions and as "secondary" rules that provide the rules of the game that is law [52, 27], emphasizing low-level decisions on concrete questions allows for people with diverse views to live together in society. Combining these pictures, we get a vision of law in which people can come to agreement on high-level questions in and underlying the law where possible but still retain the ability to cooperate where agreement is not. Specification is improved and the ability to move across levels of agreement while still creating law to address novel problems ensures that pluralism is protected.

3 Discussion

Alignment must represent the values of the society from which AI emerges, but existing approaches to incorporating democracy into alignment do not overcome either the specification or the pluralism problem. Legal theory can help. Approaches like Anthropic's Collective Constitution [23], powered by reinforcement learning from AI feedback, currently cannot specify how a model should concretely interpret the rules and principles that have been put into it; instead they "leave[] the interpretation of [the principles] to AI systems themselves" based on the models' pretraining [38]. "Good for humanity," one example of a general principle, obviously has this problem, but, as Hart demonstrated, so do all other words in rules. Simply having a group of people deliberate over what principles should be put in the model does not solve this problem and will likely lead instead to vague agreement over good-sounding concepts without a clear sense of what they mean. These existing democratic approaches similarly struggle to preserve pluralistic values. Returning to the Collective Constitution as a leading example, even where the population that performs the deliberation is representatively

---

[2]One study has shown that analogical reasoning emerges in large language models [54].

[3]Feng et al. [17] cite Sunstein in their case-based reasoning work. However, the approach here differs from their approach, which focuses on modifications of seed cases by experts rather than using cases to democratically give meaning to rules.

selected [18], the principles selected to be put into models are chosen based on the majority vote of those chosen to provide democratic input, despite substantial minorities of the voters disagreeing with the choices of the majority [23, 19]. This majoritarian selection reinforces the extent to which large language models can be understood as majoritarian machines, making completions based on the predominant perspectives in their training data and finetuning.

The legal theories discussed above can improve how democratic deliberation on alignment is done. Processes like the Collective Constitution [23] should be supplemented by the inclusion of cases and deliberation about principles of analogy to improve the degree of specification and pluralism that they allow. In one implementation of this approach, instead of merely having users write, deliberate about, and vote on abstract statements of principle, when a user selects a given statement of principle as one that they might agree with, an LLM could be used to generate a set of examples of the application of that statement in different circumstances. Users would then indicate which of the applications they feel matches their understanding of the statement, and, channeling Hart, these applications could be used to better specify how the user actually thinks about the statement that they have selected. A concept like "fairness" that is notoriously hard to define for humans, let alone computers [13, 14], becomes more tractable by leveraging contextual linguistic reasoning through cases. Next, drawing on Sunstein [50–52], the cases that users reasoned about could be compared to each other and similar decisions picked out. Data on users who selected different statements of principle but agreed to the same or similar decision of analogous concrete cases could help alignment researchers understand the underlying similarities across plural perspectives that make people decide the same way for different reasons and allow researchers to incorporate incompletely theorized agreements into models' reasoning processes. Ideally, the result of this process would be improved measures of alignment across a variety of cases. But it should also allow for more effective intercultural and global reasoning about how to align AI as well, increasing the extent to which models are pluralist.

The above proposal is merely a start, and much more work is needed to turn the idea into an implementable alignment technique. In particular, it is unclear how to get the models used to produce concrete cases to cover the whole spectrum of possible meanings of the word that the cases are intended to illustrate, especially if they are by their nature more likely to produce outputs near the most likely or core meaning of a word. Even if the approach were to be implemented and positive results demonstrated from the experiments, there would likely remain difficulties in indicating to a model how to resolve problems that actually involved disputes over higher values and rules rather than specific cases, which some leading legal theorists believe to be the core of the law [15]. Questions of structure and hierarchy that operate at the core of constitutional law would have to be resolved. Other secondary rules [27], including the extent to which different decisions became precedent that weighed against others, would have to be decided, and encouraging people to reason about those issues might be harder than thinking about direct issues of value and principle about which people tend to have intuitions. Ultimately, the best that the law has done is provide directions [52] and "rules of thumb" [15], and legal theory likely has as much to learn from alignment as vice versa. Nonetheless, this paper shows that legal theory is a rich vein upon which alignment should begin to draw and suggests both practical and theoretical avenues to do so.

4   Conclusion

AI alignment faces two key problems, pluralism and specification, that legal theory can help solve. The law has long faced the difficulty of how to have general rules that can be predictably applied in specific situations and that avoid stamping out the forms of life of those who have other values and visions of the good. Two leading legal theories suggest that the law solves these problems through the use of concrete cases that fill in the meaning of general rules in their application and that provide space for disagreement about values but alignment on outcomes. The map of possible meanings of a given term is filled out over time in a way that remains close to the actual meaning of practical cases, and people can agree on implementation while retaining different views on principle. These insights should be used to supplement existing attempts to put democratic deliberation into AI, and I have sketched out one potential approach to doing so that makes use of large language models and their capabilities to generate cases to illustrate alignment principles. However, empirical testing remains to be done. Collaboration between law and AI can help ensure that AI is aligned to diverse human values and that it becomes a force that helps people live their lives in the way best for them.

References

[1] Anthropic. Claude's constitution, 2023. url: https://www.anthropic.com/news/claudes-constitution.

[2] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. arXiv preprint, arXiv:2112.00861.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. arXiv preprint, arXiv:2212.08073.

[4] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences, 2022. arXiv preprint, arXiv:2211.15006.

[5] Amy Coney Barrett. Precedent and jurisprudential disagreement. *Texas Law Review*, 91:1711–1737, 2013.

[6] Alexander Bickel. *The Least Dangerous Branch: the Supreme Court at the Bar of Politics*. Yale University Press, 1986.

[7] Camilla Bignotti and Carolina Camassa. Legal minds, algorithmic decisions: How llms apply constitutional principles in complex scenarios, 2024. arXiv preprint, arXiv:2407.19760.

[8] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022. arXiv preprint, arXiv:2211.03540.

[9] Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable ai safety via doubly-efficient debate, 2023. arXiv preprint, arXiv:2311.14125.

[10] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. When large language models meet personalization: Perspectives of challenges and opportunities, 2023. arXiv preprint, arXiv:2307.16376.

[11] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. arXiv preprint, arXiv:1706.03741.

[12] Jan G. Deutsch. Precedent and adjudication. *Yale Law Journal*, 83(8):1553–1584, 1974.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011. arXiv preprint, arXiv:1104.3913.

[14] Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. From pseudorandomness to multi-group fairness and back, 2023. arXiv preprint, arXiv:2301.08837.

[15] Ronald Dworkin. *Law's Empire*. Belknap Press, 1988.

[16] Tyna Eloundou and Teddy Lee. Democratic inputs to ai grant program: lessons learned and implementation plans, 2024. url: https://openai.com/blog/democratic-inputs-to-ai-grant-program-update.

[17] K. J. Kevin Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X. Zhang. Case repositories: Towards case-based reasoning for ai alignment, 2023. arXiv preprint, arXiv:2311.10934.

[18] Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596:548–555, 2021.

[19] Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks, 2024. arXiv preprint, arXiv:2305.06626.

[20] Lon Fuller. Positivism and fidelity to law—a reply to professor hart. *Harvard Law Review*, 71(4):630–672, 1958.

[21] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437, 2020.

[22] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. The capacity for moral self-correction in large language models, 2023. arXiv preprint, arXiv:2302.07459.

[23] Deep Ganguli, Safron Huang, Liane Lovitt, Divya Siddarth, Thomas Liao, Amanda Askell, Yuntao Bai, Saurav Kadavath, Jackson Kernion, Cam McKinnon, Karina Nguyen, and Esin Durmus. Collective constitutional ai: Aligning a language model with public input, 2023. url: https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[24] Arthur L. Goodhart. Precedent in english and continental law. *Law Quarterly Review*, 50:40–95, 1934.

[25] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, April 2022.

[26] H.L.A. Hart. Positivism and the separation of law and morals. *Harvard Law Review*, 71(4):593–629, 1958.

[27] H.L.A. Hart. *The Concept of Law*. Oxford University Press, Oxford, 3 edition, 2012.

[28] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023. arXiv preprint, arXiv:2008.02275.

[29] Oliver Wendell Holmes, Jr. *The Common Law*. MacMillan & Co., 1881.

[30] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. arXiv preprint, arXiv:1805.00899.

[31] Phillip M. Kannan. Advisory opinions by federal courts. *University of Richmond Law Review*, 32:769–798, 1998.

[32] Atoosa Kasirzadeh and Iason Gabriel. In conversation with artificial intelligence: aligning language models with human values, 2022. arXiv preprint, arXiv:2209.00731.

[33] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents, 2021. arXiv preprint, arXiv:2103.14659.

[34] Hannah Rose Kirk, Andrew M. Bean, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and values, 2023. arXiv preprint, arXiv:2310.07629.

[35] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback, 2023. arXiv preprint, arXiv:2303.05453.

[36] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024. arXiv preprint, arXiv:2404.16019.

[37] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023. arXiv preprint, arXiv:2302.08582.

[38] Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, Ethan Perez, Jackson Kernion, Jamie Kerr, Kamal Ndousse, Karina Nguyen, Nelson Elhage, Newton Cheng, Nicholas Schiefer, Nova DasSarma, Oliver Rausch, Robin Larson, Shannon Yang, Shauna Kravec, Timothy Telleen-Lawton, Thomas I. Liao, Tom Henighan, Tristan Hume, Zac Hatfield-Dodds, Sören Mindermann, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. Specific versus general principles for constitutional ai, 2023. arXiv preprint, arXiv:2310.13798.

[39] Brian Leiter. Why legal positivism (again). *University of Chicago Public Law & Legal Theory Working Paper*, 2013.

[40] Lawrence Lessig. *Code: And Other Laws of Cyberspace*. Basic Books, 1999.

[41] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. arXiv preprint, arXiv:2212.09251.

[42] Collective Intelligence Project. Introducing the collective intelligence project, 2023. url: https://cip.org/whitepaper.

[43] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. arXiv preprint, arXiv:2305.18290.

[44] Matthew J Sanders. A california court decision—bees are fish—shines new light on the biodiversity crisis, 2022. American Bar Association.

[45] Thomas O. Sargentich. The contemporary assault on checks and balances. *Widener Journal of Public Law*, 7:231, 1998.

[46] Frederick Schauer. Precedent. *Stanford Law Review*, 39(3):571, 1987.

[47] Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback, 2022. arXiv preprint, arXiv:2204.14146.

[48] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, March 2024.

[49] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024. arXiv preprint, arXiv:2402.05070.

[50] Cass R. Sunstein. On analogical reasoning. *Harvard Law Review*, 106(3):741–791, 1993.

[51] Cass R. Sunstein. Incompletely theorized agreements. *Harvard Law Review*, 106(7):1733–1772, 1994.

[52] Cass R. Sunstein. Of artificial intelligence and legal reasoning. *University of Chicago Law School Roundtable*, 2001.

[53] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, and Tom Henighan Chris Olah. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. url: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[54] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models, 2023. arXiv preprint, arXiv:2212.09196.