Fall Semester, 2024
Instructor: J. Horty

PHIL 808m: Computational Models of Normative Reasoning
Syllabus Version #2
September 20, 2024

Description

In a number of domains—such as health-care robotics, for example—the most useful artificial autonomous agents will have to respect human norms and values. The field of artificial intelligence (AI) that is focused on designing systems capable of exhibiting this kind of normative competence is usually characterized as *machine ethics*, though a better label might be *computational normative reasoning*, since the norms involved include, not only ethical norms, but other kinds as well, such as social, legal, and cultural norms. In fact, a tremendous amount of research in AI is currently focused on the problem of designing autonomous systems exhibiting a kind of normative competence. However, this research—often carried out under the rubric of "value alignment"—is generally pursued using only a limited range of techniques, mostly centered around variants of reinforcement learning, and in almost complete isolation from research on closely related topics in ethics and legal theory. The goal of this course is to try to connect these different fields, bringing research on value alignment from AI into contact with the rich tradition of work on normative reasoning from philosophy and legal theory.

The course will be divided into three parts. First, for background and context, we will quickly review some recent work from ethical theory—on moral principles, dilemmas, particularism, and values—and then discuss two domains in which patterns of normative reasoning have been studied with some care: the common law, and bioethics. Next, we will turn to the problem of designing a system capable of acquiring, representing, and reasoning with particular human norms. After reviewing the difficulties with traditional top-down and bottom-up approaches, we will explore a new hybrid approach, based on ideas first introduced in moral and legal philosophy, and then developed in the field of AI and Law; this approach will be extended in various ways, and evaluated from analytic and experimental perspectives. Finally, we will compare this hybrid approach to some recent work in AI on value alignment and explainability.

Prerequisites: The course is interdisciplinary, and everyone will have to be willing to work through material presented in a style they may not be familiar with, whether this be purely philosophical, logical, or computational. We will assume familiarity with or willingness to learn elementary logic at the level of PHIL 271 or CMSC 250; logic at the level of PHIL 370 would be even more helpful, but not necessary. For those who might be interested in experimental work but lack necessary background, there is the possibility that a separate Python tutorial will be run in conjunction with this course.

Time and place

Wednesday, 2:30–5:00, SKN 1116

Contact information

Office: Skinner 1101. Email: horty@umd.edu. Office hours: I'll let you know my exact office hours once they've sorted themselves out. Meanwhile, please feel free to write any time if you'd like to meet.

Course materials

The readings for each topic are divided into three categories: primary readings, which you should read with some care before class sessions; secondary readings, which it would be good to read through if you can, although it is not required; and background and related material, which is provided for overall context and as a starting point for students who want to do research in a particular area. All primary readings will be on the course web site, as well as many secondary readings.

Course work

There are four kinds of course work:

Questions: Each week, each student in the class must submit at least one question concerning at least one of the primary readings assigned that week. These are due the Monday prior to class, at midnight, so that I have time to read them before class. Questions will not be graded; their purpose is simply to give me a sense of how well you're understanding the material and what your perspective on it is.

Exercises: Once we get into slightly more formal material, there will be a few simple, nuts-and-bolts, exercises. These will be assigned in class, and due at the next session. Exercises will not be graded, but you have to turn them in; the point is just to get you engaged with the material, and again, to help me track your understanding.

Presentations: The course will be run as a seminar, and everyone attending will have to help present some of the material. These presentations will be short, low-key, and likewise not graded.

Project: The main requirement for those taking the course for credit is a "project"—but since students in the class might have wildly different backgrounds and interests, I want to allow that their projects can take a variety of different forms. One option might be a standard seminar paper, of approximately 20 double-spaced pages, exploring a philosophical topic. A second option might be a more technical paper, establishing formal results about the course material. A third option might be an implementational or experimental project based on this material. A fourth option might be something else you propose that I haven't thought of—I'm open to various project ideas; and in case you have trouble isolating a project, I will develop a list of project suggestions as the term progresses.

Whatever you decided to do, we should agree on a plan by the beginning of November, so that you will have time to finish your project before end of term or shortly after. The last few class sessions will be reserved for students to present their project ideas. This, once more, is supposed to be low-key—the point is not to create stress and havoc in your lives, but simply to give you a chance to get feedback from the class and to give the class a chance to learn from your work.

Course topics

Here is a tentative, initial list of topics, which will surely be revised during the term (be sure to check the version number on the syllabus):

1. Background/overview

   (a) Machine ethics

   Primary readings: Allen et al. [6], Lazar [70], Fisher et al. [45], Moore [79]

   Seconday readings: Anderson and Anderson [7], Awad et al [10], Railton [92]

   Background, related, and additional material: Dennis et al. [41], Nallur [81], Tolmeijer et al. [122], Townsend et al. [125]

   (b) Moral reasoning

   Primary readings: Richardson [99]

   Background, related, and additional material: Cushman et al. [35], Kleiman-Weiner et al. [66]

2. Principles, particularism, conflicts, values

   (a) Principles I

   Primary readings: Hare [54, Sections 3.6–4.3], Hare [55, Chapter 3], Scanlon [109]

   Seconday readings: Scanlon [107], Scanlon [108, pp. 197–202]

   Background, related, and additional material: Schauer [111]

   (b) Particularism

   Primary readings: Dancy [37], Dancy [38], Ross [103, Chapter 2]

   Seconday readings: Dancy [36]

   Background, related, and additional material: Dancy [39], Väyrynen [129], Väyrynen [131], Wodak [133]

   (c) Moral conflicts

   Primary readings: Gowans [51, Introduction]

   Secondary readings: Brink [23], Horty [57], Pietroski [89]

   Background, related, and additional material: Connee [32], Donagan [43], Foot [46], vanFraassen [126], Hare [56, Chapter 2], Marcus [76], McConnell [78], Searle [114], Williams [132]

   (d) Principles II

   Primary readings: Richardson [97]

   Background, related, and additional material: Thakral [121], Väyrynen [130]

   (e) Values

   Primary readings: Chang [29], Chang [31],

   Secondary readings: Chang [30]

   Background, related, and additional material: Mason [77]

3. Case studies

   (a) Reasoning in the common law

   Primary readings: Alexander and Sherwin [4, Introduction, Chapters 1–3]

   Secondary readings: Raz [95, Chapter 10], Simpson [116]

   Background, related, and additional material: Alexander [3], Burton [24, Introduction and Chapters 1–4], Schauer [110], Schauer [112], Schauer [113]

   (b) Reasoning in bioethics

   Primary readings: Iltis [62],

   Background, related, and additional material: DeGrazia [40], Gert *et al.* [50], Jonsen [64], Jonsen and Toulmin [65], Little [73], Paulo [84], Paulo [85], Richardson [98], Strong [120], Toulmin [123], Toulmin [124],

4. Defeasible moral principles

   (a) Default logic

   Primary readings: Horty [58, Introduction, Chapters 1–2]

   Background, related, and additional material: Reiter [96]

   (b) A deontic interpretaton

   Primary readings: Horty [58, Chapter 3]

   Secondary readings: Maguire [75], Mullins [80]

   Background, related, and additional material: Bonevac [22], Fuhrmann [48]

5. The reason model

   (a) The model

   Primary readings: Horty [61, Introduction, Chapters 1–2], Lamond [67]

   Background, related, and additional material: Ashley [8], Eisenberg [44], Lamond [68], Levi [72, Sections I–II], MacCormick [74], Perry [86], Rissland and Ashley [101]

   (b) Dimensions

   Primary readings: Bench-Capon [13] Horty [59], Horty [60]

   Background, related, and additional material: Bench-Capon [15], Bench-Capon and Atkinson [16], Bench-Capon and Atkinson [17], Bench-Capon and Rissland [18], Rigoni [100], Rissland and Ashley [102]

   (c) Values, balancing, proportionality

   Primary readings: Bench-Capon and Sartor [19], Prakken [90], Sartor [106]

   Secondary readings: Bench-Capon [14],

   Background, related, and additional material: Alexy [5], Berman and Hafner [20], Sartor [105]

6. A hybrid approach

   (a) The basic idea

   Primary readings: Canavotto and Horty [27], Rawls [94]

   Secondary readings: Alcaraz et al. [2]

   Background, related, and additional material: Dietrich and List [42], Sher [115]

   (b) Learning a prioritized default theory

   Primary readings: [[Notes to be supplied by Jeff]]

   (c) A generalization: inconsistent case bases

   Primary readings: Canavotto [25], Canavotto [26], [[Bijan's student]]

(d) Moral databases

Readings: Bogaards [21], Sinnott-Armstrong and Skorburg [117]

Background, related, and additional material: Awad [9], Conitzer et al. [34], Freedman et al. [47], Skorburg et al. [118]

7. Value alignment and explanation

(a) Value alignment

Primary readings: Conitzer et al. [34], Gabriel [49]

Background, related, and additional material: Carrol et al. [28], Freedman et al. [47], Sorensen et al. [119]

(b) Approaches involving social choice

Primary readings: Conitzer et al. [33], [[Jeff notes, Eric/Ilaria paper]]

Secondary readings: Greene et al. [52]

Background, related, and additional material: Horty [61, Chapters 4-6], Noothigattu [82]

(c) Approaches involving LLM's

Primary readings: Jiang et al. [63], Rao et al. [93]

Secondary readings: Leike [71], Bakker et al. [12]

(d) A constitutional approach

Primary readings: Bai et al. [11]

(e) Explanation and justification

Primary readings: Lazar [69], Odederken and Bex [83], Peters et al. [87], Peters et al. [88], van Woerkom et al. [128], van Woerkom et al. [127]

Seconday readings: Prakken and Ratsma [91], Rudin [104]

Background, related, and additional material: Adadi et al. [1], Guidotti et al. [53],

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Benoit Alcaraz, Aleks Knoks, and David Streit. Estimating weights of reasons using metaheuristics: a hybrid approach to machine ethics. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES'24)*, 2024.

[3] Larry Alexander. Constrained by precedent. *Southern California Law Review*, 63:1–64, 1989.

[4] Larry Alexander and Emily Sherwin. *Demystifying Legal Reasoning.* Cambridge University Press, 2008.

[5] Robert Alexy. On balancing and subsumption: a structural comparison. *Ration Juris*, 16:433–49, 2003.

[6] Colin Allen, Iva Smit, and Wendall Wallach. Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7:149–155, 2005.

[7] Michael Anderson and Susan Leigh Anderson. Machine ethics: creating an ethical intelligent agent. *AI Magazine*, 28:15–26, 2007.

[8] Kevin Ashley. *Modeling Legal Argument: Reasoning with Cases and Hypotheticals.* The MIT Press, 1990.

[9] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-Froncois Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563:59–64, 2018.

[10] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, MJ Crockett, Jim Everett, Theodoros Evgeniou, Alison Gopnik, Julian Jamison, Tae Wan Kim, Matthew Liao, Michael Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, and Juliani Schroeder. Computational ethics. *Trends in Cognitive Science*, 26:388–405, 2022.

[11] 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.

[12] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael H. 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. Unpublished manuscript, available at https://arxiv.org/abs/2211.15006, 2022.

[13] Trevor Bench-Capon. Some observations on modelling case based reasoning with formal argument models. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law (ICAIL-99)*, pages 36–42. The Association for Computing Machinery Press, 1999.

[14] Trevor Bench-Capon. The missing link revisited: the role of teleology in representing legal argument. *Artificial Intelligence and Law*, 10:79–94, 2002.

[15] Trevor Bench-Capon. Representing Popov v. Hayashi with dimensions and factors. *Artificial Intelligence and Law*, 20:15–35, 2012.

[16] Trevor Bench-Capon and Katie Atkinson. Dimensions and values for legal CBR. In *Proceedings of the Thirtieth International Conference on Legal Knowledge and Information Systems (JURIX 2017)*, pages 27–32. IOS Press, 2017.

[17] Trevor Bench-Capon and Katie Atkinson. Lessons from implementing factors with magnitude. In *Proceedings of the Thirty First International Conference on Legal Knowledge and Information Systems (JURIX 2018)*, pages 11–20. IOS Press, 2018.

[18] Trevor Bench-Capon and Edwina Rissland. Back to the future: dimensions revisited. In *The Fourteenth Annual Conference on Legal Knowledge and Information Systems (JURIX-2001)*, pages 41–52. IOS Press, 2001.

[19] Trevor Bench-Capon and Giovanni Sartor. Theory based explanation of case law domains. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Law (ICAIL-2001)*, pages 12–21. The Association for Computing Machinery Press, 2001.

[20] Donald Berman and Carole Hafner. Representing teleological structure in case-based legal reasoning: the missing link. In *Proceedings of the Fourth International Conference on Artificial Intelligence and Law (ICAIL-93)*, pages 50–59. The Association for Computing Machinery Press, 1993.

[21] Ellen Bogaards. Comparing case-base classification models with interpretable machine learning classification models: a study based on human moral judgement in the bioethical domain. Masters Thesis, Utrecht University, 2023.

[22] Daniel Bonevac. Defaulting on reasons. *Nous*, 52:229–259, 2016.

[23] David Brink. Moral conflict and its structure. *The Philosophical Review*, 103:215–247, 1994.

[24] Steven Burton. *An Introduction to Law and Legal Reasoning.* Little, Brown, and Company, 1985.

[25] Ilaria Canavotto. Precedential constraint derived from inconsistent case bases. In *Proceedings of the Thirty-fifth Annual Conference on Legal Knowledge and Information Systems (JURIX 2022)*, pages 23–32, 2022.

[26] Ilaria Canavotto. Reasoning with inconsistent precedents. *Artificial Intelligence and Law*, forthcoming.

[27] Ilaria Canavotto and John Horty. Piecemeal knowledge acquisition for computational normative reasoning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22)*, pages 171–180, 2022.

[28] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235. PMLR, 2024.

[29] Ruth Chang. Value incomparability and incommensurability. In Iwao Hirose and Jonas Olson, editors, *The Oxford Handbook on Value Theory*, pages 205–224. Oxford University Press, 2015.

[30] Ruth Chang. Comparativism: the grounds of rational choice. In Barry McGuire and Errol Lord, editors, *Weighing Reasons*, pages 213–240. Oxford University Press, 2016.

[31] Ruth Chang. Parity: an intuitive case. *Ratio*, pages 196–411, 2016.

[32] Earl Conee. Against moral dilemmas. *Philosophical Review*, 91:87–97, 1982.

[33] Vincent Conitzer, Rachel Freeman, Jobst Hertzig, Wesley Holliday, Bob Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William Zwicker. Social choice should guide AI alighment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[34] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First National Conference on Artificial Intelligence (AAAI-17)*, pages 4831–4835, 2017.

[35] Fiery Cushman, Victor Kumar, and Peter Railton. Moral learning: psychological and philosophical perspectives. *Cognition*, 167:1–10, 2017.

[36] Jonathan Dancy. Can particularists learn the difference between right and wrong? In *Proceedings of the Twentieth World Congress of Philosophy*, pages 59–72. Philosophy Documentation Center, 1999.

[37] Jonathan Dancy. Intention and permissibility. In *Proceedings of the Aristotelian Society, Supplementary Volume 74*, pages 319–338, 2000.

[38] Jonathan Dancy. Moral particularism. In Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy (Summer 2001 Edition)*. Stanford University, 2001. Available at http://plato.stanford.edu/archives/sum2001/entries/moral-particularism/.

[39] Jonathan Dancy. *Ethics Without Principles*. Oxford University Press, 2004.

[40] David DeGrazia. Moving forward in bioethical theory: theories, cases, and specified principlism. *Journal of Medicine and Philosophy*, 17:511–539, 1992.

[41] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[42] Franz Dietrich and Christian List. A reason-based theory of rational choice. *Nous*, 47:104–134, 2013.

[43] Alan Donagan. Consistency in rationalist moral systems. *The Journal of Philosophy*, 81:291–309, 1984.

[44] Melvin Eisenberg. *The Nature of the Common Law*. Harvard University Press, 1988.

[45] Michael Fisher, Christian List, Marija Slavkovik, and Alan Winfield. Dagstuhl manifesto: Engineering moral machines. *Informatik Spektrum*, 39:467–489, 2016.

[46] Philippa Foot. Moral realism and moral dilemmas. *Journal of Philosophy*, 80:379–398, 1983.

[47] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:1–14, 2020.

[48] André Fuhrmann. Extensions and projections in deontic default logic. Unpublished manuscript, 2017.

[49] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

[50] Bernard Gert, Charles Culver, and K. Danner Clouser. Common morality versus specified principlism: reply to Richardson. *Journal of Medicine and Philosophy*, 25:308–322, 2000.

[51] Christopher Gowans, editor. *Moral Dilemmas*. Oxford University Press, 1987.

[52] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. Embedding ethical principles in collective decision support systems. In *Proceedings of the Thirtieth National Conference on Artificial Intelligence (AAAI-16)*, pages 4147–4151, 2016.

[53] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.

[54] R. M. Hare. *The Language of Morals.* Oxford University Press, 1952.

[55] R. M. Hare. *Freedom and Reason.* Oxford University Press, 1963.

[56] R. M. Hare. *Moral Thinking: Its Levels, Method, and Point.* Oxford University Press, 1980.

[57] John Horty. Reasoning with moral conflicts. *Nous*, 37:557–605, 2003.

[58] John Horty. *Reasons as Defaults.* Oxford University Press, 2012.

[59] John Horty. Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, 27:307–345, 2019.

[60] John Horty. Modifying the reason model. *Artificial Intelligence and Law*, 29:271–285, 2021.

[61] John Horty. *The Logic of Precedent: Constraint, Freedom, and Common Law Reasoning.* Cambridge University Press, in press.

[62] Ana Smith Iltis. Bioethics as methodological case resolution: specification, specified principlism, and casuistry. *Journal of Medicine and Philosophy*, 25:271–284, 2000.

[63] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *CoRR*, abs/2110.07574, 2021.

[64] Albert Jonsen. Strong on specification. *Journal of Medicine and Philosophy*, 25:348–360, 2000.

[65] Albert Jonsen and Stephen Toulmin. *The Abuse of Casuistry: A History of Moral Reasoning.* University of California Press, 1988.

[66] Max Kleiman-Weiner, Rebecca Saxe, and Joshua Tenenbaum. Learning a commonsense moral theory. *Cognition*, 167:107–123, 2017.

[67] Grant Lamond. Do precedents create rules? *Legal Theory*, 11:1–26, 2005.

[68] Grant Lamond. Precedent and analogy in legal reasoning. In Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)*. Stanford University, 2006. Available at http://plato.stanford.edu/archives/spr2016/entries/legal-reas-prec/.

[69] Seth Lazar. Legitimacy, authority, and democratic duties of explanation. Unpublished manuscript, forthcoming in *Oxford Studies in Political Philosophy*, 2022.

[70] Seth Lazar. Frontier AI ethics: anticipating and evaluating the societal impacts of generative agents. Unpublished manuscript, 2024.

[71] Jan Leike. A proposal for importing society's values: building towards Coherent Extrapolated Volition with language models. *Musings on the Alignment problem*, 2023. Available at https://aligned.substack.com/p/a-proposal-for-importing-societys-values.

[72] Edward Levi. *An Introduction to Legal Reasoning.* The University of Chicago Press, 1949.

[73] Margaret Little. Moral generalities revisited. In Brad Hooker and Margaret Little, editors, *Moral Particularism.* Oxford University Press, 2000.

[74] Neil MacCormick. Why cases have *rationes* and what these are. In Laurence Goldstein, editor, *Precedent in Law*, pages 155–182. Oxford University Press, 1987.

[75] Barry Maguire. The value-based theory of reasons. *Ergo*, 3:233–262, 2016.

[76] Ruth Barcan Marcus. Moral dilemmas and consistency. *Journal of Philosophy*, 77:121–136, 1980.

[77] Elinor Mason. Value pluralism. In Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy (Summer 2023 Edition)*. Stanford University, 2023. Available at https://plato.stanford.edu/archives/sum2023/entries/value-pluralism/.

[78] Terrance McConnell. Moral dilemmas and consistency in ethics. *Canadian Journal of Philosophy*, pages 269–287, 1978.

[79] James H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[80] Robert Mullins. Formalizing reasons, oughts, and requirements. *Ergo an Open Access Journal of Philosophy*, 7, 2021.

[81] Vivek Nallur. Landscape of machine implemented ethics. *Science and Engineering Ethics*, 26:2381–2399, 2020.

[82] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making, 2018.

[83] Daphne Odekerken and Floris Bex. Towards transparent human-in-the-loop classification of fraudulent web shops. In *The Thirty-Third Annual Conference on Legal Knowledge and Information Systems (JURIX-2020)*, volume 334 of *Frontiers in Artificial Intelligence and Applications*, pages 239–242. IOS Press, 2020.

[84] Norbert Paulo. *Methods in applied ethics: a view from legal theory.* PhD thesis, Philosophy Department, University of Hamburg, 2014.

[85] Norbert Paulo. Specifying specification. *Kennedy Institute of Ethics Journal*, 26:1–28, 2016.

[86] Stephen Perry. Judicial obligation, precedent, and the common law. *Oxford Journal of Legal Studies*, 7:215–257, 1987.

[87] Joeri Peters, Floris Bex, and Henry Prakken. Model- and data-agnostic justifications with *a fortiori* case-based argumentation. In *Proceedings of the Nineteenth Internation Conference on Artificial Intelligence and Law (ICAIL-23)*, pages 207–216. The Association for Computing Machinery Press, 2023.

[88] Joeri Peters, Henry Prakken, and Floris Bex. Justification derived from inconsistent case bases using authoritativeness. In *Proceedings of the First International Workshop on Argumentation for eXplainable AI (ArgXAI)*, volume 3209. CEUR Workshop Proceedings, 2022.

[89] Paul Pietroski. Prima facie obligations, ceteris paribus laws in moral theory. *Ethics*, 103:489–515, 1993.

[90] Henry Prakken. An exercise in formalising teleological case-based reasoning. *Artificial Intelligence and Law*, 10:113–133, 2002.

[91] Henry Prakken and Rosa Ratsma. A top-level model of case-based argumentation for explanation: Formalisation and experiments. *Argument & Computation*, 13(2):159–194, 2022.

[92] Peter Railton. Ethical learning, natural and artificial. In S. Matthew Liao, editor, *Ethics of Artificial Intelligence*, pages 45–78. Oxford University Press, 2020.

[93] Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159. Association for Computational Linguistics, 2023.

[94] John Rawls. Outline of a decision procedure for ethics. *Philosophical Review*, 60:177–197, 1951.

[95] Joseph Raz. *The Authority of Law.* Oxford University Press, 1979.

[96] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[97] Henry Richardson. Specifying norms as a way to resolve concrete ethical problems. *Philosophy and Public Affairs*, 19:279–310, 1990.

[98] Henry Richardson. Specifying, balancing, and interpreting bioethical principles. *Journal of Medicine and Philosophy*, 25:285–307, 2000.

[99] Henry Richardson. Moral reasoning. In Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Stanford University, 2018.

[100] Adam Rigoni. Representing dimensions within the reason model of precedent. *Artificial Intelligence and Law*, 26:1–22, 2018.

[101] Edwina Rissland and Kevin Ashley. A case-based system for trade secrets law. In *Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL-87)*, pages 60–66. The Association for Computing Machinery Press, 1987.

[102] Edwina Rissland and Kevin Ashley. A note on dimensions and factors. *Artificial Intelligence and Law*, 10:65–77, 2002.

[103] W. D. Ross. *The Right and the Good.* Oxford University Press, 1930.

[104] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[105] Giovanni Sartor. Doing justice to rights and values: teleological reasoning and proportionality. *Artificial Intelligence and Law*, 18:175–215, 2010.

[106] Giovanni Sartor. Consistency in balancing: from value assessments to factor-based rules. In D. Duarte and S Sampaio, editors, *Proportionality in Law: An Analytical Perspective*, pages 121–136. Springer, 2018.

[107] T. M. Scanlon. Contractualism and utilitarianism. In Amartya Sen and Bernard Williams, editors, *Utilitarianism and Beyond*, pages 103–1283. Cambridge University Press, 1982.

[108] T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.

[109] T. M. Scanlon. Intention and permissibility. In *Proceedings of the Aristotelian Society, Supplementary Volume 74*, pages 301–317, 2000.

[110] Frederick Schauer. Formalism. *The Yale Law Journal*, 97:509–548, 1988.

[111] Frederick Schauer. Exceptions. *The University of Chicago Law Review*, 58:871–899, 1991.

[112] Frederick Schauer. *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and Life*. Oxford University Press, 1991.

[113] Frederick Schauer. *Thinking Like a Lawyer*. Harvard University Press, 2009.

[114] John Searle. Prima-facie obligations. In Zak van Straaten, editor, *Philosophical Subjects: Essays Presented to P. F. Strawson*, pages 238–259. Oxford University Press, 1980.

[115] Itai Sher. Comparative value and the weight of reasons. *Economics and Philosophy*, pages 103–158, 2019.

[116] A. W. B. Simpson. The *ratio decidendi* of a case and the doctrine of binding precedent. In A. G. Guest, editor, *Oxford Essays in Jurisprudence*, pages 148–175. Oxford University Press, 1961.

[117] Walter Sinnot-Armstrong and Joshua August Skorburg. How AI can aid bioethics. *Journal of Practical Ethics*, 9, 2021.

[118] Joshua Skorburg, Walter Sinnott-Armstrong, and Vincent Conitzer. AI methods in bioethics. *AJOB Empirical Bioethics*, 11:37–39, 2020.

[119] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher M. Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235. PMLR, 2024.

[120] Carson Strong. Specified principlism: what is it, and does it really resolve cases better than casuistry? *Journal of Medicine and Philosophy*, 25:323–341, 2000.

[121] Ravi Thakral. Moral principles as generics. *Journal of the American Philosophical Association*, 10:1–20, 2023.

[122] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: A survey. *ACM Computing Surveys*, 53(6), 2021.

[123] Stephen Toulmin. The tyranny of principles. *The Hastings Center Report*, 11:31–39, 1981.

[124] Stephen Toulmin. How medicine saved the life of ethics. *Perspectives in Biology and Medicine*, 25:736–750, 1982.

[125] Beverley Townsend, Colin Paterson, T. T. Arvind, Gabriel Nemirovsky, Radu Calinescu, Ana Cavalcanti, Ibrahim Habli, and Alan Thomas. From pluralistic normative principles to autonomous-agent rules. *Minds and Machines*, 32:683–715, 2022.

[126] Bas van Fraassen. Values and the heart's command. *The Journal of Philosophy*, 70:5–19, 1973.

[127] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. Justification in case-based reasoning. In *Proceedings of the First International Workshop on Argumentation for eXplainable AI (ArgXAI)*, volume 3209. CEUR Workshop Proceedings, 2022.

[128] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. Landmarks in case-based reasoning: from theory to data. In Stefan Schlobach, Maria Perez-Ortiz, and Myrthe Tielman, editors, *Proceedings of the First International Confefence on Hybrid Human-Machine Intelligence*, pages 212–224. IOS Press, 2022.

[129] Pekka Väyrynen. Moral particularism. In Christian B. Miller, editor, *Continuum Companion to Ethics*, pages 247–260. Continuum Books, 2011.

[130] Pekka Väyrynen. Reasons and moral principles. In Daniel Star, editor, *The Oxford Handbook of Reasons and Normativity*, pages 839–861. Oxford University Press, 2018.

[131] Pekka Väyrynen. Moral generalism and moral particularism. In Christian B. Miller, editor, *Bloomsbury Handbook of Ethics*, pages 381–396. Bloomsbury, 2023.

[132] Bernard Williams. Ethical consistency. *Proceedings of the Aristotelian Society*, 39 (supplemental):103–124, 1965. A revised version appears in *Problems of the Self: Philosophical Papers 1956–1972*, Cambridge University Press, 1973, pages 166–186.

[133] Daniel Wodak. Moral perception, inference, and intuition. *Philosophical Studies*, 4176:1495–1512, 2019.