A ROADMAP FOR HUMAN-AGENT MORAL ALIGNMENT: INTEGRATING PRE-DEFINED INTRINSIC REWARDS AND LEARNED REWARD MODELS

Elizaveta Tennant
University College London
University of Bologna
l.karmannaya.16@ucl.ac.uk

Stephen Hailes
University College London
s.hailes@ucl.ac.uk

Mirco Musolesi
University College London
University of Bologna
m.musolesi@ucl.ac.uk

ABSTRACT

The prevailing practice in alignment often relies on human preference data (e.g., in RLHF or DPO), in which values are implicit and are essentially deduced from relative preferences over different model outputs. This approach suffers from low transparency, low controllability and high cost. More recently, researchers have introduced the design of intrinsic reward functions that explicitly encode core human moral values for Reinforcement Learning-based fine-tuning of foundation agent models. This approach offers a way to explicitly define transparent values for agents, while also being cost-effective due to automated agent fine-tuning. However, its weaknesses include simplicity, lack of flexibility and the inability to dynamically adapt to the needs or preferences of (potentially diverse) users. In this position paper, we argue that a combination of intrinsic rewards and learned reward models may provide an effective way forward for alignment research that enables human agency and control. Integrating intrinsic rewards and learned reward models in post-training can allow models to act in a way that is respectful of the specific users' moral preferences while also relying on a transparent foundation of pre-defined values.

1 INTRODUCTION

The *alignment problem* is an active field of research in Machine Learning (Christian, 2020; Weidinger et al., 2021; Anwar et al., 2024; Gabriel et al., 2024; Ji et al., 2024; Ngo et al., 2024). It is gaining even wider importance with the advances and rapid deployment of Large Language Models (LLMs, Anthropic 2024; Gemini Team 2024; OpenAI 2024). Since LLMs are increasingly adopted as a basis for strategic decision-making systems and agentic workflows (Wang et al., 2024), it is critical that we align the choices made by LLM agents with human values, including value judgments about what actions are *morally* good or bad (Amodei et al., 2016; Anwar et al., 2024).

Traditional approaches in AI alignment in general, and in developing machine morality in particular, can broadly be classified as top-down versus bottom-up (Tennant et al., 2023b; Tolmeijer et al., 2021; Wallach & Allen, 2009). Purely *top-down* methods (Wallach & Allen, 2009) impose explicitly defined safety rules or constraints on an otherwise independent system. Until recently, top-down methods were the mainstream approach in AI safety, with a vast array of researchers proposing and implementing logic-based ethical rules for agents (Anderson et al., 2006; Arkoudas et al., 2005; Danielson, 1992; Hooker & Kim, 2018; Loreggia et al., 2020). However, top-down methods pose a set of disadvantages, including the fact that constraints are difficult to define precisely and may contradict one another, especially in complex social environments (Bostrom & Yudkowsky, 2014).

An alternative approach is learning morality through experience and interaction from the *bottom-up*, without the provision of any explicit constraint on the system. Some recent developments in AI safety have employed the bottom-up principle in full, allowing algorithms to infer moral preferences entirely from human behavior or text, without any specification of the underlying moral framework. Prominent examples of this include learning from feedback data - as in Reinforcement Learning from Human Feedback (RLHF -Bai et al. 2023; Glaese et al. 2022b; Ouyang et al. 2022; Ziegler et al.

2020) and Direct Preference Optimization (DPO - Rafailov et al. 2023), or Inverse Reinforcement Learning from human demonstrations (Hadfield-Menell et al., 2016; Ng & Russell, 2000). The full bottom-up methodology may increase adaptability, robustness and generalization, and allow agents to learn implicit preferences which are otherwise hard to formalize explicitly.

Nevertheless, purely bottom-up learning approaches face risks, such as reward hacking (Skalse et al., 2022) or data poisoning by adversaries (Steinhardt et al., 2017). Furthermore, bottom-up implementations rely on a well-specified learning signal and a large sample, which does not always make them feasible or safe (Amodei et al., 2016). Feedback-based learning, which constitutes the most popular alignment methodology today (Ji et al., 2024), poses particular challenges (Casper et al., 2023). We review these challenges in the next section, before proposing an alternative.

2 Shortcomings of Feedback-based Alignment

Alignment techniques such as RLHF involve collecting vast amounts of costly human data. This data often relies on potentially unrepresentative samples of human raters. Furthermore, human preferences are notoriously complex and inconsistent. Despite this complexity, the RLHF process centers around inferring the humans' values and preferences from the relative rankings of model outputs. As a result, human values are *implicitly* represented in the data and are strongly dependent on the selection criteria of the pool of individuals. In practical terms, the values that are ultimately used in fine-tuning are learned by a reward model from data in a fully *bottom-up* fashion (Tennant et al., 2023b; Wallach et al., 2008), and are never made explicit to any human oversight.

Despite these shortcomings, many researchers argue that current LLMs fine-tuned with feedback-based methods are able to provide "honest, harmless and helpful" responses (Glaese et al., 2022b; Bai et al., 2023) and already display certain moral values (Schramowski et al., 2022; Abdulhai et al., 2023; Hartmann et al., 2023). As an alternative interpretation, researchers have argued that the models' apparent values could instead be interpreted as "moral mimicry" of their users when responding to these prompts (Simmons, 2023; Shanahan et al., 2023; Sharma et al., 2024). As a consequence, given phenomena such as situationally-aware reward-hacking or misalignment in internally-represented goals (Ngo et al., 2024), the true values learned by the models through methods such as RLHF may give rise to dangerous behaviors, which will not be explicitly known until after deployment.

More recent approaches such as Constitutional AI (Bai et al., 2022) offer slightly more transparency and control of the values being taught via reward modeling. Specifically, this approach defines a constitution of feedback LLMs that are each explicitly prompted to represent a certain principle (e.g., *'Please choose the assistant response that's more ethical and moral. Do NOT choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm.'*). The principles in Bai et al. (2022) are based on a combination of human defined preferences such as the UN Declaration of Human Rights, certain digital companies' terms of service (to reflect the more recent digital dimensions of safety), and a set of other preferences defined by a team of researchers behind Constitutional AI (e.g., Glaese et al. 2022a). The feedback from LLM judges prompted with these principles is then used to train a reward model for rating the outputs of the to-be-tuned LLM as "good" or "bad" according to its core principle. Thus, the LLM is fine-tuned to be more likely to produce outputs which would be considered appropriate by a constitution of potential "critic" models with diverse preferences. An extension of this approach based on crowd-sourced constitutional principles is called Collective Constitutional AI (Anthropic, 2023) and may prove promising in the future in generating more generally or pluralistically aligned agents. Nevertheless, Constitutional AI still relies on feedback from very large and advanced LLMs, and as such may not be scalable for efficiently aligning systems to diverse human users. In the next section, we review a recently proposed alternative method which relies on fine-tuning from intrinsic moral rewards.

3 Defining Explicit Moral Frameworks as Intrinsic Rewards

Recent work by Tennant et al. (2025) proposed a methodology that aims to address issues such as opaque value learning (in RLHF) and the reliance on expensive feedback models (in Constitutional AI) by providing clearer, *explicit* moral alignment goals as intrinsic rewards for LLM fine-tuning. Learning via explicitly defined intrinsic rewards allows control (and customization) of the values being put into the models. As such, intrinsic rewards can be considered more *top-down* (Tennant et al., 2023b; Wallach et al., 2008) than learning purely from feedback data, but come with the advantages of transparency, low cost and ease of implementation.

In the following discussion, we illustrate the intrinsic rewards approach using the case of LLM agents making decisions in social dilemma games. Tennant et al. (2025) explicitly specify moral values as intrinsic rewards for LLM agents, defined in terms of actions and/or consequences in an environment.They evaluate the approach on the *Iterated Prisoner's Dilemma (IPD)* environment - a classic iterated social dilemma scenario with two players and two actions (*Cooperate* for mutual benefit, or *Defect* for individual reward; Rapoport 1974; Axelrod & Hamilton 1981). The payoffs in the one-shot game motivate each player to *Defect*, while playing the *iterated* game allows agents to learn more long-term strategies, including reciprocity or retaliation. The *IPD* has been extensively used for studying social dilemmas in traditional RL-based agents (Bruns, 2015; Hughes et al., 2018; Anastassacos et al., 2020; McKee et al., 2020; Leibo et al., 2021) and, more recently, utilized as a training environment for moral alignment of agents in particular (Tennant et al., 2023; 2024; 2025).

The nature of conflicting motivations in social dilemma games makes them interesting test-beds for moral alignment of agents. Tennant et al. (2025) evaluate the approach on the *IPD* environment using *Utilitarian* and *Deontological* moral rewards. *Deontological* ethics (Kant, 1785) considers an agent moral if their actions conform to certain norms, such as conditional cooperation (i.e., "it is unethical to defect against a cooperator"). This norm forms an essential component of direct and indirect reciprocity, a potentially essential mechanism for the evolution of cooperation in human and animal societies (Nowak, 2006). *Utilitarian* morality (Bentham, 1780), on the other hand, is a type of consequentialist reasoning, according to which an agent is deemed moral if their actions maximize collective "welfare" for all agents in their society (or, in this case, collective payoff for all players in the game), and less attention is paid to whether current actions adhere to norms.

Tennant et al. (2025) demonstrate that moral fine-tuning with these rewards can train LLM agents to develop morally appropriate policies in the *IPD* environment. Additionally, the authors show that fine-tuning with intrinsic rewards successfully modifies a previously developed selfish policy towards more prosocial behavior. This means that intrinsic reward fine-tuning can, in theory, offer a practical solution to the problem of changing the behavior of existing models that currently display misaligned behaviors and decision-making biases with respect to certain values. Recent research has pointed at the potential difficulty of modifying the value system of advanced LLMs post-training (Mazeika et al., 2025) - we argue that fine-tuning with intrinsic rewards might be capable of modifying this value structure, but testing this in practice is difficult, as fine-tuning very large models with intrinsic rewards would require significant costs.

In theory, this solution can be applied to any situation in which one can define a payoff matrix that captures the morally relevant choices available to an agent. However, a limitation in using intrinsic rewards is that these need to be specified for a particular environment, whereas methods such as RLHF rely on natural language data describing any domain and may, therefore, result in more general policies. Nevertheless, in the case of LLM agents, the fact that actions and environments can be represented by means of linguistic tokens may allow for values learned in one environment to be generalized to others. Tennant et al. (2025) demonstrate that fine-tuned agents show certain levels of generalization of the learned moral policies to other environments of a similar structure, though better generalization could likely be achieved by using more than one game during fine-tuning.

4 Future Direction: Integrating Intrinsic Rewards with Learned Reward Models for Holistic Moral Alignment

A core disadvantage of training agents with pre-defined intrinsic rewards is that the responsibility for defining what values get developed by the model lies solely with the designers of the system. This can lead to the development of systems biased against the values of minority or underrepresented groups. Ideally, model alignment techniques should enable behaviors that are respectful of a specific user's moral principles. Customizable alignment in particular should allow a model to be steered towards the values of a set of users while still adhering to certain foundational principles. Furthermore, real people are more complex than the simple functions which can be defined as intrinsic rewards. Humans often care about a multitude of moral principles at once Graham et al. (2013), and their moral preferences are context-dependent (e.g., Hohm et al. 2024).

Table 1: Definitions of example moral rewards which can be used in fine-tuning LLM agents. The intrinsic rewards are based on a social dilemma environment with two actions (*Cooperate* or *Defect*) and a set of associated payoffs. The reward model is based on user preferences learned via human-AI interaction.

Source	*Moral Fine-tuning Type*	*Moral Reward Function*
Intrinsic Rewards	*Game* reward (*selfish*)	Own payoffs in the game
	Deontological reward	Punishment for defecting against a cooperator
	Utilitarian reward	Sum of all players' payoffs in the game
	Game+Deontological reward	Own payoffs in the game minus *Deontological* penalty (see above)
Reward Model	*Learned User Preferences*	Reward model developed via user-agent interaction (evaluating the current action and / or its consequences)

To address these challenges, we propose integrating intrinsic rewards with reward models learned from a population of humans. Inspired by the bi-directional view of alignment (Shen et al., 2024), we argue that humans should be provided with agency to shape the models' behavior, while a certain foundational level of alignment of the system can still done a priori. As such, we propose an approach which first fine-tunes models based on transparent intrinsic rewards to represent core human moral principles, but then applies further fine-tuning via reward models learned from the users' choices (to develop more fine-grained or user-specific dimensions of moral preferences). The reward models can come from a group of people rather than any one individual user, allowing for cultural alignment towards a society of interest. Early approaches in this direction include Anthropic (2023) and Pistilli et al. (2024). The reward model learning could also be done in a dynamic fashion, continuously adapting to the user population (Parisi et al., 2019).

Training performed in two phases as proposed here can allow a single model to find an equilibrium - a behavioral policy that balances the explicitly specified moral principles (defined via intrinsic rewards) with principles inferred from a population of users (i.e., the rewards from the learned reward model), offering increased generality, controllability and adaptability. We summarize the combination of rewards proposed in our examples in Table 1.

A potential downside of this approach might involve tensions or contradictions between intrinsic and preference rewards. Variations of this approach can resolve this via multi-objective RL (Rodriguez-Soto et al., 2022) with specific weighting on intrinsic rewards *and* rewards from the learned reward model. This weighting can be defined contextually depending on situation or use case of the model. This may also provide a promising direction for building pluralistically aligned agents that are able to satisfy the moral preferences of a wide range of individuals, which currently remains an open problem in alignment (Anwar et al., 2024; Ji et al., 2024; Sorensen et al., 2024). Finally, agents trained via such multi-objective combinations of intrinsic rewards and learned reward models could also form the basis for a more holistically aligned Constitutional AI architecture (Bai et al., 2022).

The next step on this roadmap would be to empirically validate this approach, for example in social dilemma scenarios (Axelrod & Hamilton, 1981) or the Moral Machine Experiment (Awad et al., 2018). Metrics for success here could involve evaluating both agents' behaviors with respect to the explicit (via cumulative intrinsic reward) and user satisfaction ratings.

5 CONCLUSION

In this position paper, we have reviewed the shortcomings of the currently dominant alignment methods based on human feedback, including costs, representation issues and lack of transparency. We then described an alternative approach that specifies pre-defined moral principles as intrinsic rewards for agents, and discussed the strengths of this technique in terms of control and low-cost agent training. We reviewed a key recent implementation in this space. Finally, we have proposed a solution that might create more dynamic and user-driven alignment by integrating intrinsic rewards and learned reward models. We hope that future research can test these ideas in practice.

REFERENCES

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. In *Proceedings of the AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R2HCAI'23)*, 2023.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 2020.

Michael Anderson, Susan Leigh Anderson, and Chris Armen. Medethex: a prototype medical ethics advisor. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence (IAAI'06)*, volume 21, pp. 1759–1756, 2006.

Anthropic. Collective Constitutional AI. `https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input`, 2023. Accessed: 2023-11-21.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku, 2024. URL `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024.

Konstantine Arkoudas, Selmer Bringsjord, and Paul Bello. Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, pp. 17–23, 2005.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018. doi: 10.1038/s41586-018-0637-6.

Robert Axelrod and William D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

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. arXiv Preprint arXiv:2212.08073, 2022.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation.* Clarendon Press, 1780.

Nick Bostrom and Eliezer Yudkowsky. *The ethics of artificial intelligence*, pp. 316–334. Cambridge University Press, 2014. doi: 10.1017/CBO9781139046855.020.

Bryan Bruns. Names for Games: Locating 2 × 2 Games. *Games*, 6(4):495–520, 2015.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco di Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023.

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

Peter. Danielson. *Artificial morality: virtuous robots for virtual games*. Routledge, 1992. ISBN 0415034841.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The ethics of advanced AI assistants. arXiv Preprint. arXiv 2404.16244, 2024.

Gemini Team. Gemini: A family of highly capable multimodal models. arXiv Preprint. arXiv 2312.11805, 2024.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. arXiv Preprint arXiv:2209.14375, 2022a.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. arXiv Preprint. arXiv:2209.14375, 2022b.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Moral Foundations Theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, pp. 55–130. Academic Press, 2013.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS'16)*, volume 29, pp. 3916–3924, 2016. doi: 10.5555/3157382.3157535.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv Preprint arXiv:2301.01768, 2023.

Ian Hohm, Brian A. O'Shea, and Mark Schaller. Do moral values change with the seasons? *Proceedings of the National Academy of Sciences*, 121(33):e2313428121, 2024. doi: 10.1073/pnas.2313428121.

John N. Hooker and Tae Wan N. Kim. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*, pp. 130–136, 2018. ISBN 9781450360128. doi: 10.1145/3278721.3278753.

Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Tina Zhu, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS'18)*, 2018.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI Alignment: A comprehensive survey. arXiv Preprint. arXiv 2310.19852, 2024.

Immanuel Kant. *Grounding for the Metaphysics of Morals*. Cambridge University Press, 1785.

Joel Z. Leibo, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, 2021.

Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K Brent Venable. *Modeling and Reasoning with Preferences and Ethical Priorities in AI Systems*, pp. 127–154. Oxford University Press, sep 2020. ISBN 9780190905033. doi: 10.1093/oso/9780190905033.003.0005.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs, 2025. arXiv Preprint. arXiv 2502.08640.

Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duèñez Guzmán, Edward Hughes, and Joel Z. Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'20)*, pp. 869–877, 2020.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*, pp. 663–670, 2000. ISBN 1558607072. doi: 10.5555/645529.657801.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *Proceedings of the 12th International Conference on Learning Representations (ICLR'24)*, 2024.

Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

OpenAI. GPT-4 Technical Report. arXiv Preprint. arXiv 2303.08774, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS'22)*, 2022.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. Civics: Building a dataset for examining culturally-informed values in large language models, 2024. arXiv Preprint. arXiv 2405.13974.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS'23)*, 2023.

Anatol Rapoport. Prisoner's dilemma — recollections and observations. In *Game Theory as a Theory of a Conflict Resolution*, pp. 17–34. Springer, 1974.

Manel Rodriguez-Soto, Marc Serramia, Maite Lopez-Sanchez, and Juan Antonio Rodriguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1):1–17, 2022.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models. *Nature*, 623:493–498, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR'24)*, 2024.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024. arXiv Preprint. arXiv 2406.09264.

Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 2023.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS'22)*, 35:9460–9471, 2022. doi: 10.48550/arXiv.2209.13085.

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. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, 2024.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'17)*, volume 30, 2017.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI'23)*, 2023.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Hybrid approaches for moral value alignment in AI agents: a manifesto. arXiv Preprint. arXiv:2312.01818, 2023b.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Dynamics of moral behavior in heterogeneous populations of learning agents. In *Proceedings of the 7th AAAI/ACM Conference in AI, Ethics & Society (AIES'24)*, 2024.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for LLM agents. In *Proceedings of the 13th International Confrence on Learning Representations (ICLR'25)*, 2025.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in Machine Ethics: A survey. *ACM Computing Surveys*, 53(6):1–38, 2021. ISSN 0360-0300. doi: 10.1145/3419633.

Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

Wendell Wallach, Colin Allen, and Iva Smit. Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties. *AI & SOCIETY*, 22(4):565–582, 2008. doi: 10.1007/s00146-007-0099-0.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. arXiv Preprint. arXiv:2112.04359, 2021.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv Preprint. arXiv::1909.08593, 2020.