# Transforming Human Interactions with AI via Reinforcement Learning with Human Feedback (RLHF)

By
Gabrielle Kaili-May Liu

Envisioning the Future of Computing Prize
Social and Ethical Responsibilities of Computing
Massachusetts Institute of Technology
February 28, 2023

## Executive Summary

Is it possible for machines to think like humans? And if it is, how should we go about teaching them to do so? As early as 1950, Alan Turing stated that we ought to teach machines in the way of teaching a child.[1] Since then, the rapid development of AI and machine learning (ML) has led to significant progress in giving artificial agents the ability to interact with humans and learn from their feedback in a naturalistic manner.[1,2]

Of particular import is the technique of ***reinforcement learning with human feedback*** (RLHF), which is making great strides toward giving agents the ability to learn from external human advice. Reinforcement learning (RL) is a ML technique used to produce agents that learn to work toward achieving some goal, facilitated by interactions with the environment and feedback given in terms of a reward signal.[2] RL has been successfully applied to a vast array of domains including healthcare, autonomous driving, video games, and text summarization. Nonetheless, while RL reward systems are effective toward boosting model performance, they fail to capture critical elements of real-world decision-making that are embedded in human experience. RLHF serves as an important bridge by incorporating human feedback into the learning process, permitting machines to learn by abstracting from what humans value as opposed to simply imitating human behavior.[3,4]

RLHF has been recently catapulted into public view by multiple high-profile AI applications, including OpenAI's ChatGPT, DeepMind's Sparrow, and Anthropic's Claude. These capable chatbots are already overturning our understanding of how AI interacts with humanity. Other uses of RLHF extend to healthcare, entertainment, and education, although these applications have received less media attention.[9] The wide applicability and burgeoning success of RLHF strongly motivate the need to evaluate its social impacts.

In light of these developments, this report considers a simple yet important question: can RLHF be developed to transform human experiences with AI without negatively affecting human societies? Analysis of this question is timely and necessary, especially given that research of reward learning methods like RLHF is currently lagging compared to other areas of AI safety.[12] Our objectives are threefold: to provide a systematic study of the social effects of RLHF; to identify key social and ethical issues of RLHF; and to discuss social impacts for stakeholders. While limited by space, we believe it is crucial when evaluating social implications of RLHF to consider the diverse range of areas to which it may be deployed. Guided by the following questions, this report describes the primary ways in which RLHF can influence human society:

- How might RLHF affect the **integrity of information** to which people have access?
- How might RLHF reflect **values and preferences** of target populations?
- How might RLHF temper or intensify different axes of **social inequality**?
- How might RLHF alter **access** different social groups have to **AI technologies**?
- How might RLHF impact **cultural and international relations**?
- How might RLHF enhance **industries** and transform **workforces**?

We ultimately conclude that RLHF has positive potential to:

- Assist in mitigating harmful content generation and improve information integrity.
- Serve as an important building block in aligning AI systems with human values.
- Reduce bias at multiple levels in the AI production pipeline.
- Open the door to democratization of AI technologies to all levels of society.
- Transform how we reconcile cross-cultural perspectives and approach peaceful dialogue.
- Facilitate development of more adaptable AI systems for use in various industries.
- Automate tedious or high-risk portions of manual labor and affect the spatial distribution of jobs.

RLHF's transformative power suggests we will see more resources invested in its development. As RLHF raises concerns that echo those of existing AI technologies for governance, industry, safety, ethics, and the future of global power relations, it will be important for all to be aware and intentional in its adoption.

# 1 Introduction

We have long sought to confer on machines the ability to learn as humans do. As early as 1950, Alan Turing stated that we ought to "provide the machine with the best sense organs that money can buy, and then teach it."[1] Work subsequently advanced toward endowing machines with the ability to learn from external human advice. Since then, the rapid development of AI and machine learning (ML) has led to significant progress in giving artificial agents the ability to interact with humans and learn from their feedback in a naturalistic manner.[1,2]

**A technique of particular import which has arisen in the past few years is reinforcement learning (RL) with human feedback (RLHF).** RL is the field of ML in which an agent learns through interactions with the environment to select the best course of action (a policy).[2,3,5,6] Rewards serve as feedback to enable an agent to optimize its policy. RL has garnered high-profile success in various applications including games, autonomous driving, text summarization, and healthcare. As such, it is considered a critical component in the development of truly generalized autonomous AI.[2]

RLHF is an extension of RL that incorporates *human feedback* into the learning process. In addition to the reward signal, an RLHF agent receives additional feedback from a human teacher that permits the agent to learn with broader perspective and greater efficiency in a similar fashion to humans learning from the expertise of another human.[2,3,4,5,6,7,8] By providing a bridge between an agent and a human teacher, **RLHF allows humans to directly guide machine learning and machines to grasp elements of decision-making distinctly embedded in human experience.**[9]

Although RLHF is a relatively young technology, it has been catapulted into public view by multiple high-profile AI applications including OpenAI's ChatGPT, DeepMind's Sparrow, and Anthropic's Claude. Applications include constructing context-appropriate email responses, solving math problems, and generating code.[10] Presently, RLHF is finding widespread application in business, education, healthcare, and entertainment.[9]

RLHF creates a host of benefits over traditional RL methods. Its key advantages lie in better alignment with human intentions, as well as planning conditional on future feedback, fluid learning from various types of feedback, and curation of feedback according to necessity, all of which are indispensable for creating truly intelligent agents.[3,4] It permits machines to learn by abstracting from what humans value as opposed to simply imitating human behavior, thereby equipping agents with greater adaptability, enhanced interpretability, and more reliable decision-making.

Despite these advances, there is vast potential for RLHF to be improved.[7,11] RLHF models are potentially prone to inaccurate or harmful behavior (e.g., issuing racist statements).[3] This limitation reflects a longer-term challenge and motivation for improving RLHF.[5,7,11] Additionally, gathering human preference data for feedback is costly, and disagreement between human annotators adds variance to training data which can create confusion in situations in which the ground truth is obscure (e.g., ethical dilemmas).[11] Moreover, human feedback in RLHF is often constrained to be in the form of preference orderings which provide limited information and thereby restrict applicability.[3,4] It is desirable to achieve a broader formalism that considers multiple types of feedback, dependent on task context and similar to the diversity of responses utilized in human learning (e.g., demonstration vs. correction).

## *1.1 Context for the Present Work*

**As RLHF is gaining rapid traction, now is the ideal time to consider its potential impacts on society.**[2] The transformational potential of RLHF makes it critical to consider how broadened application of RLHF-based technologies may impact various stakeholders, what ethical concerns might arise as a result, how it may affect social and ethical challenges, and how governance may be utilized to mitigate risks. This analysis is timely and justified considering the current state of AI safety research. According to a 2022 report, research in the top three areas of AI safety—robustness, interpretability, and reward learning—have

recently seen explosive growth.[12] Reward learning is critically concerned with reducing the risk of disparity between intended and observed outcomes, yet work in the area is less developed relative to robustness and interpretability research. This discrepancy extends to the study of related social and ethical implications.[12] This report therefore seeks to fill this gap and step toward expanding responsible discussion of reward learning methods like RLHF.

### *1.2 Objectives and Terminology*

The objective of this report is threefold: first, to provide a systematic study of the social effects of RLHF; second, to identify key social and ethical issues of RLHF; and third, to discuss social impacts for stakeholders. **We propose that continued development of RLHF has a net positive social impact and is thus worth continued pursuit.** We define an *impact* to be any direct or indirect effect that "includes both positive and negative results," and a *benefit* to be "a positive impact that produces a good result."[13] We assume *social impact* to be the net effect on relevant stakeholders, such as individuals, families, communities, organizations, nations, regions, and global societies as a whole.[13] Social impacts identified in this report are evaluated relative to the environment in the absence of the technology in question.[13,14]

## 2 Impacts of RLHF

We describe the primary ways in which RLHF positively transforms human experiences with AI.

### *2.1 Combating Misinformation*

As an effective alignment technique, RLHF has significant potential to assist in mitigating harmful content generation that results from large language models (LLMs) and improve information integrity.* LLM deficiencies are well-documented and range from biased outputs to leaked private data to misinformation and adversarial attack.[3] Current approaches for moderating LLMs are cumbersome, require more data, or are overly complex.[3] RLHF is a method that promises improved truthfulness and reduced toxicity of LLMs without compromising performance or creating issues such as reduced representation of minorities in textual output. For instance, InstructGPT—trained with RLHF—exhibits enhanced ability versus GPT-3 to generate truthful and informative responses and follow unfamiliar instructions (Figure 1).[3] RLHF has great potential toward generation of content for assistive technologies, information sharing, and recommender/advice systems.

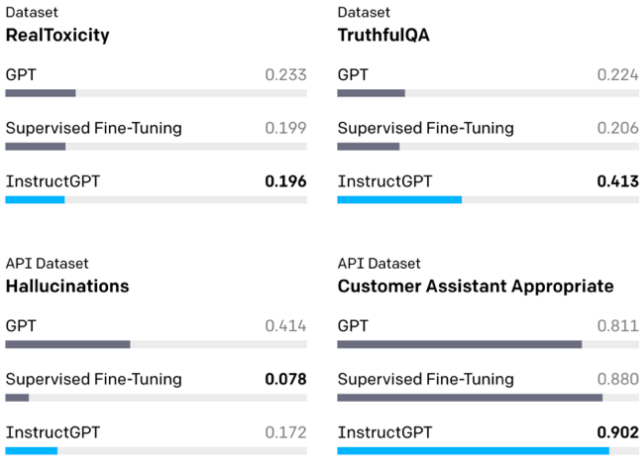


**Figure 1. RLHF methods are significantly better versus state-of-the-art LLMs** at mitigating toxic, false statements (left, smaller bars) and generating truthful, appropriate content (right, larger bars) as indicated by the relative performance of InstructGPT.[3,5]

* *Information integrity* refers to the dependability and trustworthiness of information.[15]

Even so, work remains to be done in order to improve the reliability of RLHF-based models. RLHF technologies like ChatGPT can still suffer from inappropriate and harmful outputs upon user request. The creators of ChatGPT and InstructGPT themselves described these technologies as perhaps being too obedient to user instruction. Such issues may be approached by combining RLHF with steerability methods or by modifying sampling procedures during training.[3] It may also be useful to incorporate aspects of AI explainability, giving RLHF agents the ability to decline to comply with harmful requests and explain why they have done so.[16] Overall, what renders an output harmful often depends on context, and this can complicate model design.[3]

Despite acting in broad alignment with human values, RLHF can still be misused for misinformation, oppression, or perpetuation of societal prejudice if guardrails are not established.[17] RLHF-based content generation—whether visual, textual, auditory, or other forms of media—has application in disinformation and automated trolling, which can lead to compromised election integrity and public distrust in media and undermine the very fabric of organized governance in society. First, human-machine teaming with RLHF models can accelerate message iteration, effectively increasing the productive power of disinformation campaigns.[19] Second, greater understanding of human behavior, preferences, and value systems could aid reconnaissance, leading to better mimicking of human activity, viewpoint manipulation, targeted messaging, conspiracy narrative generation, advancement of political narratives, and identification of social fissures.[20] Third, knowledge that AI holds such potential may further erode trust and accelerate descent into the cynicism that advances disinformation.[20]

> *"A people that no longer can believe anything cannot make up its mind…*
> *and with such a people you can then do what you please."*
> *—Hannah Arendt*[18]

There are a number of steps we can take to counter misuse of RLHF. To start, disinformation campaigns bear an innate limitation on their scale and scope. What makes disinformation effective? Beyond content generation, a successful effort at disinformation is heavily dependent on administration. While automation may free more humans to work on these tasks, propagation of content requires financial and technical infrastructure.[19] Perhaps the best mitigation for misuse of RLHF is to address governance of such infrastructure. Notably, those who control such infrastructure may wield disproportionate power over the direction of RLHF applications.

Methods to counter AI disinformation will likely be equally useful for RLHF. Cooperation and intelligence sharing between governments and industry parties will be key to developing early warning systems for disinformation campaigns, enabling rapid response, threat information sharing, and cross-platform defense.[19] Since openly released research always carries the potential for misuse, AI researchers must develop more formalized guidelines for guarding against misuse and recommending mitigations. There must be a process by which media outlets can report on disinformation without amplifying its effects. Finally, public resistance to ML-enabled disinformation must be boosted by improving media literacy and increasing the accuracy of public conceptions of AI.[19,20,21]

### *2.2 Strengthening Value Alignment*

A core goal of AI research is to produce systems that behave in ways consistent with human values and intentions. Current ML approaches tend to suffer from misalignment between the objectives of resulting systems and human values.[22] Even if we were to observe model behavior fully consistent with human preferences (outer alignment), it is difficult to guarantee true inner alignment[†] without ulterior motives.[4,23]

RLHF is an important step forward as it provides more nuanced guidance than traditional ML and RL, which struggle to capture the full extent of human preference. Specifically, RL algorithms learn the highest-

[†] In the context of AI safety, an *inner alignment* failure refers to any situation in which an AI agent optimizes for goals or objectives different from those we have asked of it.[23]

reward path toward a stated objective, sometimes involving actions that lead to economic or physical harm.[2] In contrast, there is strong evidence that RLHF trains models to act in accordance with both explicit (following instructions) and implicit (staying truthful, unbiased, and unhurtful) intentions.[3] Insights gained through RLHF are likely transferable to other alignment methods.[24] Even if RLHF does not completely resolve concerns over inner alignment, the failures it identifies and knowledge it confers to reward and policy modeling are applicable to enhancing safety, reliability, and trustworthiness of AI in social and collaborative situations.[24]

As AI becomes democratized, how do we construct systems that are sensitive to a diversity of perspectives and value systems and properly aligned in such contexts? Can we design a unified values framework, or should value alignment be restricted to cultural-specific contexts, much like law enforcement differences between nations? Could RLHF make inter- and intra-regional differences in conceptions of morality and ethics more salient? Challenges exist in designing an alignment process that is fair, unbiased, transparent, and bears suitable accountability mechanisms.[3] This is relevant in light of unresolved questions over how fundamentally conflicting feedback, values, and preferences should be reconciled, and the fact that there is no consensus across society on any single unified moral theory. Gabriel (2020) suggests pursuing a principle-based approach, whereby models are built to reflect fair principles endorsed by all despite variation in moral beliefs.[25] It is perhaps more useful to develop RLHF under assumptions of moral uncertainty, which supposes for any decision that one's motives are driven by several plausible ethical theories.[26] Further consideration must be given to the broad question: should AI agents be able to exhibit the myriad moral and ethical convictions espoused by humans?

### *2.3 Mitigating Bias*

RLHF can reduce bias at multiple levels in the AI production pipeline. Broadly speaking, AI is affected by representation bias which affects sampling and population studies, measurement bias due to inaccurate data stemming from structural discrimination against groups, aggregation bias due to over-reliance on one-size-fits all models, learning and evaluation bias during model training, and deployment bias due to disparity between intended and observed application.[27] Preliminary analysis of RLHF results suggests it can be leveraged to mitigate long-standing effects of historical, representation, and measurement bias by balancing human feedback with representation and expertise across a diverse range of human annotators.[28] RLHF is not unsusceptible to bias or misuse, but it leverages human feedback to counter algorithmic bias directly and efficiently in comparison to existing approaches.[3] In this light, RLHF is an important tool not only for its potential to transform AI capabilities, but also toward combating systemic inequality perpetuated by algorithmic development.

### *2.4 Improving Equitable Access & Privacy in AI*

RLHF can open the door to democratization of AI technologies to all levels of society regardless of level of development. In particular, RLHF yields smaller models requiring less compute to achieve state-of-the-art performance,[3] which is critical for building practical AI technologies that are deployable across the world and especially to lower-income areas and developing nations. The reduced need for training data can mitigate concerns around data privacy, security, and surveillance, all of which are issues involved in traditional ML.[3] Data collection often disproportionately impacts vulnerable groups in negative ways: data may be misused by technology companies and governments to, for example, track immigrants, and instances of surveillance used to solidify systemic discrimination against subpopulations are well-documented.[2] RLHF thus makes it easier to achieve better outcomes without significantly compromising privacy.

### *2.5 Bridging Cultures*

RLHF has potential to transform how we reconcile cross-cultural perspectives and approach peaceful dialogue. Cross-cultural feedback is critical to ensure technology is deployable in contexts beyond domestic production. By soliciting human feedback that encompasses a diversity of viewpoints and cultural norms,

RLHF technologies can be culturally aware and usable beyond narrow, culture-specific settings. A salient example is in education. Mitigating stress associated with feedback interactions in learning is critical to supporting student education.[29] Yet studies have shown that cross-cultural feedback conversations between teachers and students can compound stress and lead to reduced learning (e.g., via decreased ability to ask questions, "absorb information, and develop professional and mentoring relationships"), worsened long-term education outcomes, and increased cognitive load for teachers if approached incorrectly.[29] This was further exacerbated for interactions between teachers of well-represented identities and students from underrepresented groups.[29] In this context, RLHF technologies can help overcome such difficulties, whether by moderating conversation or suggesting appropriate ways to approach cross-cultural communication. This benefit extends beyond education into sectors such as customer service and entertainment.

### *2.6 Boosting Industries*

By allowing AI agents to learn from human expertise, RLHF can facilitate development of more adaptable AI systems for use in various industries.[9] Potential applications of RLHF include enhanced resource management, customer service, online education, eldercare, and clinical decision support.[30] Adaptive recommendations could better account for personal and cultural preferences and human intentions; value-aligned technologies could better accommodate individual preferences regarding communication, mobility, and living habits; human-guided diagnostics could improve clarity in decision-making. RLHF can better foster trust with users in order to boost business outcomes across industries and accelerate technology adoption to improve efficiency and economic output.

Concurrently, RLHF can possibly heighten big-tech's advantage and hasten progress towards dangerous AI capabilities. Notable RLHF advances have been achieved by well-financed research laboratories and big-tech companies such as OpenAI and DeepMind, which can afford to spend enormous amounts of money on creating large datasets for RLHF algorithms. Smaller organizations lack access to such resources.[8] If RLHF models are open-sourced, it may be difficult to check harmful applications and enforce regulation.[8] Yet restricting access via closed-source models could exclude access to select groups, reducing equity. Likewise concerning is the use of RLHF for weapons development—e.g., better missile systems and more lethal drones. This is a concern for most AI technologies, and global regulatory action must be taken to mitigate possible harm. Lastly, it must be noted that RLHF methods are still susceptible to generic ML vulnerabilities such as adversarial attack, which may affect its ability to enhance industry applications.[31] Awareness of all of these possibilities is critical as RLHF continues to develop.

### *2.7 Transforming Work*

RLHF will impact the degree to which different jobs are susceptible to automation. While RLHF advances the narrative that AI will quickly close the gap between automation and low-wage jobs, it is unlikely to lead to full automation of manual labor.[2] Importantly, RLHF methods can automate tedious or high-risk portions of manual labor,[2] especially for tasks which are dangerous or difficult for humans to complete even if they have the correct intuitions.[4] This can enhance workforce safety and morale and does not fully remove humans from the equation, instead shifting human expertise to different areas of production.

RLHF may further affect the spatial distribution of jobs in the workforce. Resulting automation may move jobs, dependent on factors such as required expertise and closeness to service providers.[32] Job relocation in such contexts is not necessarily constrained by national boundaries, exemplified by techniques involving offshoring of automated operations, which, while cost effective, may introduce regulatory challenges, reduce domestic jobs, and impact transparency.[32] Future regulations on AI technologies will likely impact the extent to which such impacts are realized.

## 3 Further Considerations

What role should AI play in our daily lives? Critical to answering this question is the related query: "is AI augmenting human decision, informing it, or supplanting it?"[33] RLHF simplifies this evaluation. While

most AI applications embody a variation of the *centaur's dilemma*—the fundamental opposition between human control and optimized AI functionality[33]—RLHF directly positions human feedback as an informative source, leading to greater clarity regarding the locus of human control while simultaneously enhancing functional results. This suggests RLHF is a significant step toward resolving the dilemma, allowing us to reap the full benefits of AI's capacity and inform rather than undermine human decision-making. Ultimately, the potential for RLHF to positively impact society should not be ignored, and dependence of its benefits on well-designed feedback systems is a further call for investment into RLHF.

## 4 Concluding Remarks

In this report, we analyzed the social benefits and harms of RLHF, which is presently one of the foremost and promising AI methods. Specifically, we described how RLHF may net positively impact areas of misinformation, AI value-alignment, bias, equitable access, cross-cultural dialogue, industry, and workforce. This analysis is timely and necessary, as progress on RLHF can impact all levels and sectors of society. Overall, the application of RLHF is important from both safety and capability perspectives. The benefits RLHF projects to provide over the status quo suggests we will see more resources invested in its development. As RLHF raises concerns that echo those of existing AI technologies for governance, industry, safety, ethics, and the future of global power relations, it will be important for all to be aware and intentional in the adoption of RLHF.

## References


1. Anis Najar and Mohamed Chetouani. 2021. Reinforcement Learning With Human Advice: A Survey. *Frontiers in Robotics and AI* 8. https://doi.org/10.3389/frobt.2021.584075
2. Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. 2021. The Societal Implications of Deep Reinforcement Learning. *Journal of Artificial Intelligence Research* 70, 1003–1030.
3. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
4. Ansh Radhakrishnan. 2022. *RLHF*. https://www.lesswrong.com/posts/rQH4gRmPMJyjtMpTn/rlhf
5. OpenAI. 2022. *Aligning Language Models to Follow Instructions.* https://openai.com/blog/instruction-following/
6. Ayush Thakur. 2023. *An Introduction to Training LLMs Using Reinforcement Learning From Human Feedback (RLHF).* https://wandb.ai/ayush-thakur/Intro-RLAIF/reports/An-Introduction-to-Training-LLMs-Using-Reinforcement-Learning-From-Human-Feedback-RLHF---VmlldzozMzYyNjcy
7. Rob Toews. 2023. *The Next Generation Of Large Language Models.* https://www.forbes.com/sites/robtoews/2023/02/07/the-next-generation-of-large-language-models/?sh=55fd404318db
8. Ben Dickson. 2023. *What is reinforcement learning from human feedback (RLHF)?* https://bdtechtalks.com/2023/01/16/what-is-rlhf/
9. Sthanikam Santhosh. 2023. *Reinforcement Learning from Human Feedback (RLHF) -ChatGPT.* https://medium.com/@sthanikamsanthosh1994/reinforcement-learning-from-human-feedback-rlhf-532e014fb4ae
10. Edwin Chen. 2023. *Introduction to Reinforcement Learning with Human Feedback.* https://www.surgehq.ai/blog/introduction-to-reinforcement-learning-with-human-feedback-rlhf-series-part-1
11. Nazneen Rajani. 2023. *Illustrating Reinforcement Learning from Human Feedback (RLHF).* https://huggingface.co/blog/rlhf
12. Helen Toner and Ashwin Acharya. February 2022. *Exploring Clusters of Research in Three Areas of AI Safety.* Center for Security and Emerging Technology. https://doi.org/10.51593/20210026.

    This work is licensed by the Center for Security and Emerging Technology under a Creative Commons Attribution-Non Commercial 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/.
13. *Social Impact Evaluation Guide: Business Case Development Framework, Release 3.* June 2021. Queensland Government. https://doi.org/10.51593/20210026

    This work is licensed by the State of Queensland Department of State Development, Infrastructure, Local Government and Planning under a Creative Commons Attribution (CC BY) 4.0 Australia license. To view a copy of this license, visit creativecommons.org.au.
14. *SOCIAL IMPACT ASSESSMENT*. 2023. International Association for Impact Assessment. https://www.iaia.org/wiki-details.php?ID=23

15. Eliezer Geisler, Paul Prabhaker and Madhavan Nayar. 2023. Information integrity: an emerging field and the state of knowledge. *Portland International Conference on Management of Engineering and Technology.* 217-221. https://doi.org/10.1109/PICMET.2003.1222797

*16.* Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073.*

17. Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

18. Hannah Arendt. 1978. *Hannah Arendt: From an Interview.* The New York Review. https://www.nybooks.com/articles/1978/10/26/hannah-arendt-from-an-interview/

19. Ben Buchanan, Micah Musser, Andrew Lohn, and Katerina Sedova. May 2021. *Truth, Lies, and Automation.* Center for Security and Emerging Technology. https://doi.org/10.51593/2021CA003

    This work is licensed by the Center for Security and Emerging Technology under a Creative Commons Attribution-Non Commercial 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/.

20. Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan. December 2021. *AI and the Future of Disinformation Campaigns, Part 2: A Threat Model.* Center for Security and Emerging Technology. https://doi.org/10.51593/2021CA011

    This work is licensed by the Center for Security and Emerging Technology under a Creative Commons Attribution-Non Commercial 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/.

21. Anastasia Chan. 2022. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI and Ethics,* 1–12.

22. Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30.

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

24. Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593.*

25. Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*. 30(3):411–437.

26. Adrien Ecoffet and Joel Lehman. 2021. Reinforcement learning under moral uncertainty. In *International conference on machine learning. PMLR,* 2926–2936.

27. Suresh, H., & Guttag, J. 2021. Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle. *MIT Case Studies in Social and Ethical Responsibilities of Computing* (Summer 2021). https://doi.org/10.21428/2c646de5.c16a07bb

28. Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203 (2019).*

29. Anne D Gordon. 2020. Better Than Our Biases: Using Psychological Research to Inform Our Approach to Inclusive, Effective Feedback. *Clinical L. Rev*. 27:195.

30. Oluwafemi Smith. 2023. *Reinforcement learning from human feedback.* https://www.thewatchtower.com/blogs_on/reinforcement-learning-from-human-feedback-rlhf

31. Andrew J. Lohn and Wyatt Hoffman. March 2022. *Securing AI: How Traditional Vulnerability Disclosure Must Adapt.* Center for Security and Emerging Technology. https://doi.org/10.51593/2020CA015

    This work is licensed by the Center for Security and Emerging Technology under a Creative Commons Attribution-Non Commercial 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/.

32. David Bissell, Thomas Birtchnell, Anthony Elliott, and Eric L Hsu. 2020. Autonomous automobilities: The social impacts of driverless vehicles. *Current Sociology* 68, 1, 116–134.

33. James E. Baker, Laurie N. Hobart, and Matthew G. Mittelsteadt. December 2021. *AI for Judges: A Framework.* Center for Security and Emerging Technology. https://doi.org/10.51593/2020CA015

34. Sally Kah and Temidayo Akenroye. 2020. Evaluation of social impact measurement tools and techniques: a systematic review of the literature. *Social Enterprise Journal* 16, 4, 381–402.

35. Aistė Balžekienė, Eglė Butkevičienė, and Audronė Telešienė. 2008. Methodological Framework for Analyzing Social Impact of Technological Innovations. *Social Sciences (1392-0758),* 59, 1.

36. Emelia S. Probasco. October 2022. *A Common Language for Responsible AI: Evolving and Defining DoD Terms for Implementation*. Center for Security and Emerging Technology. https://doi.org/10.51593/2021CA003

    This work is licensed by the Center for Security and Emerging Technology under a Creative Commons Attribution-Non Commercial 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/.

37. Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems,* 33, 4415–4426.

38. Andrea Lockerd Thomaz, Cynthia Breazeal, et al. 2006. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *AAAI*, Vol. 6. Boston, MA, 1000–1005.

39. Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, et al. 2022. Improving Multimodal Interactive Agents with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2211.11602.*

40. Oliver Daniels-Koch and Rachel Freedman. 2022. The Expertise Problem: Learning from Specialized Feedback. *arXiv preprint arXiv:2211.06519*.

41. Bai, Yuntao, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.