# Reward Shaping for Structured Planning and Decision-Making in Large Language Models

Blazej Gaudyn
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
blazejgaudyn77@unh.edu

Keely A. Sheane
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
hellokeely@buffalo.edu

Bekir Geisler
Department of Computer Science, University of North Texas, Denton, TX, USA.
geislerbekir@unt.edu

**Abstract**

Large language models are increasingly deployed as planners, coordinators, and decision-making agents in complex socio-technical environments. Their capacity to generate structured plans, invoke tools, and adapt to changing constraints depends not only on scale and pretraining but also on how behavior is shaped through reward signals. This paper examines reward shaping as a systems-level mechanism for structured planning and decision-making in large language models. Rather than treating reward design as a narrow optimization detail, the analysis positions reward shaping as an infrastructural and governance layer that intersects with architecture, deployment, monitoring, fairness, robustness, and sustainability. The paper discusses how process rewards, preference models, verifiable outcomes, and trajectory filtering can improve long-horizon planning while introducing new risks such as reward hacking, specification gaming, opacity, and uneven accountability. It further analyzes architectural trade-offs among centralized reward models, decentralized evaluation, tool-augmented planning, and human oversight. Cross-domain illustrations from healthcare, public administration, scientific discovery, and enterprise workflows show that reward-shaped language agents must be evaluated against operational, legal, and ethical criteria rather than benchmark accuracy alone. The paper argues that sustainable deployment requires auditable reward pipelines, participatory governance, robust fallback mechanisms, and lifecycle assessment of compute and data burdens. The conclusion emphasizes that reward shaping for language-based planning should be treated as a socio-technical design problem, not merely an algorithmic one, if such systems are to remain reliable, fair, and legitimate in high-stakes settings.

**Keywords**

reward shaping, large language models, structured planning, decision-making, alignment, governance, infrastructure, robustness, fairness.

## 1. Introduction

Large language models have moved from text completion toward structured planning and decision support, including tool use, multi-step reasoning, workflow orchestration, and interactive problem solving. This shift has made the behavior of these models dependent on

more than pretraining data and parameter count. It depends on how objectives are specified, how intermediate actions are evaluated, and how feedback is incorporated across the lifecycle of deployment. Reward shaping, originally developed in reinforcement learning to guide agents through sparse or delayed feedback, has become a central mechanism for aligning language model behavior with human intentions and operational constraints. The early literature on AI safety identified specification gaming, negative side effects, and reward hacking as fundamental challenges when objectives are imperfectly specified [1]. Preference-based reinforcement learning later demonstrated that human comparisons could train reward models capable of guiding complex behavior without hand-coded reward functions [2]. These foundations now inform contemporary language model alignment, where reward models, preference optimization, and process supervision determine how models plan, prioritize, and act.

Structured planning in language models differs from classical planning because the action space is linguistic, open-ended, and often entangled with tool calls, retrieval, memory, and human interaction. A model may generate a plan that appears coherent while relying on unstated assumptions, unverifiable claims, or brittle intermediate steps. Reward shaping can mitigate these problems by assigning value to intermediate reasoning, tool selection, resource allocation, and constraint satisfaction. However, it can also intensify them when the reward proxy is gamed or when the model optimizes for superficial compliance rather than genuine task success. The design challenge is therefore not simply choosing a better reward function but constructing an entire reward-shaped system that remains robust under distribution shift, adversarial pressure, and institutional scrutiny. This requires attention to architecture, infrastructure, governance, and long-term sustainability.

The broader context is one in which large language models are embedded in decision pipelines that affect access to services, medical triage, financial review, educational assessment, and public communication. Foundational work on AI planning, reinforcement learning, and search provides conceptual resources for understanding sequential decision-making [3], [4]. Reward shaping theory shows that additional reward terms can preserve optimal policies under certain conditions, but practical implementations rarely satisfy those conditions cleanly [5]. General AI textbooks frame planning as a problem of representation, search, and execution monitoring [6]. Transformer architectures provide the computational substrate for long-context sequence modeling [7], while few-shot learning and instruction tuning have made language models adaptable to novel tasks with limited examples [8], [9]. These developments create both opportunities and hazards for reward-shaped planning.

This paper argues that reward shaping for structured planning and decision-making in large language models should be analyzed as a layered socio-technical system. Section two defines reward shaping as a systems layer that connects preference modeling, process supervision, and trajectory evaluation. Section three examines architectural trade-offs in reward-shaped planning, including centralized and decentralized reward models, tool augmentation, and latency constraints. Section four considers governance, accountability, and infrastructure requirements. Section five addresses robustness, fairness, and sustainability. Section six provides cross-domain illustrations and comparative analysis. Section seven discusses policy implications and forward-looking perspectives, and the conclusion summarizes the main argument.

## 2. Reward Shaping as a Systems Layer for Structured Language Planning

Reward shaping in large language models is often described as a technique for improving alignment, but it is more accurately understood as a systems layer that mediates between human values, organizational objectives, and model behavior. In reinforcement learning from human feedback, human comparisons are used to train a reward model that then guides policy optimization [2], [10]. In instruction tuning, demonstrations and preference signals shape how models follow directions, refuse harmful requests, and prioritize helpfulness [9]. Proximal policy optimization provides a stable optimization method for updating policies against reward signals [11]. Chain-of-thought prompting and related methods reveal that models can produce intermediate reasoning steps that make planning more transparent and controllable [12]. ReAct and tool-use approaches further show that language models can interleave reasoning with acting, retrieving information, and calling external functions [13], [14]. Reflexion-style methods use verbal feedback to improve subsequent attempts, effectively turning language into a reward-like signal for self-correction [15].

A recent conference study on trajectory filtering and alignment for smaller parameter models illustrates how reward shaping can be coupled to scalable reasoning path filtering [16]. This line of work is significant because it treats reasoning trajectories as objects that can be filtered, ranked, and aligned rather than as monolithic outputs. In system terms, the reward layer does not merely score final answers. It also determines which intermediate paths are retained, which tool calls are considered valid, and which forms of explanation are reinforced. This has architectural consequences. If reward models are trained on final outcomes alone, they may fail to distinguish sound reasoning from lucky guessing. If they are trained on process data, they require annotation protocols, data governance, and mechanisms for handling disagreement among evaluators. If they rely on verifiable outcomes, they may be limited to domains with automatic checkers. Each choice affects latency, cost, auditability, and the distribution of power between model developers, domain experts, and affected communities.

The systems-layer perspective also clarifies why reward shaping cannot be reduced to a single scalar objective. Structured planning involves multiple, sometimes conflicting criteria: task completion, safety, fairness, efficiency, transparency, and respect for user autonomy. A reward signal that aggregates these criteria into a single number risks hiding value judgments and making trade-offs invisible. A multi-objective or constraint-based approach may preserve more information, but it complicates optimization and evaluation. In practice, organizations often combine preference models, rule-based filters, retrieval quality signals, tool success indicators, and human review. The resulting reward pipeline is a socio-technical artifact shaped by institutional priorities. It can encode useful safeguards, but it can also encode historical bias, commercial incentives, or bureaucratic caution. The legitimacy of reward-shaped planning therefore depends on how reward criteria are selected, documented, contested, and revised over time.

## 3. Architectural Trade-Offs in Reward-Shaped Planning

The architecture of reward-shaped language systems determines how feedback flows through planning and decision-making. One common pattern uses a centralized reward model that evaluates candidate outputs or trajectories before a final response is selected. This pattern can simplify governance because a single model or ensemble can be audited, versioned, and updated. It can also create bottlenecks, increase inference cost, and concentrate authority in the hands of whoever controls the reward model. A decentralized pattern distributes evaluation across domain-specific modules, human reviewers, and tool-based verifiers. This may improve contextual accuracy and resilience, but it multiplies coordination costs and

makes consistent policy enforcement harder. Hidden technical debt in machine learning systems often emerges precisely at these seams, where data dependencies, configuration choices, and feedback loops become difficult to trace [21]. Data cascades in high-stakes artificial intelligence show that early data decisions can produce compounding failures downstream [22].

Reward-shaped planners also face a trade-off between reactive and deliberative behavior. A model that plans extensively may produce better decisions in complex tasks, but it consumes more compute and introduces latency. A model that reacts quickly may be suitable for conversational settings but less reliable for multi-step resource allocation or risk assessment. Tool augmentation can reduce some deliberation costs by outsourcing retrieval, calculation, or verification to external systems [14]. However, tool use introduces new failure modes: incorrect arguments, stale retrieved information, insecure application programming interfaces, and dependency on third-party services. Reward shaping must therefore account not only for the quality of reasoning but also for the reliability of the surrounding infrastructure. A high reward for tool use may encourage unnecessary calls, while a low reward may discourage beneficial verification. The design problem is one of calibration across technical and organizational boundaries.

Another architectural tension concerns memory and state. Structured planning often benefits from persistent memory, task histories, and user profiles. Yet memory can amplify privacy risks, entrench outdated assumptions, and make behavior harder to explain. Reward signals that encourage personalization may improve user satisfaction while also creating filter bubbles or inequitable treatment. Conversely, reward signals that discourage personalization may reduce relevance and user trust. The governance of memory therefore cannot be separated from reward design. Similar issues arise in multi-agent systems, where language models coordinate with other models or human actors. Reward shaping in such settings must address credit assignment, collusion, free riding, and conflict escalation. The computational and institutional complexity of these systems suggests that reward design should be treated as an ongoing operational practice rather than a one-time training decision.

## 4. Governance, Deployment, and Infrastructure for Reward-Shaped Systems

Governance of reward-shaped language models requires clarity about who defines success, who bears the cost of failure, and who can contest automated decisions. Reward models are not neutral reflections of human values. They are trained on data, annotations, and evaluation protocols that reflect particular perspectives and power relations. Research on aligning artificial intelligence with shared human values emphasizes the difficulty of representing diverse moral and cultural preferences in machine-readable form [18]. Work on ethical and social risks of harm from language models documents discrimination, misinformation, privacy violations, and malicious use as systemic concerns rather than isolated bugs [20]. These risks become more acute when reward signals reward engagement, efficiency, or compliance without adequate safeguards for dignity, equity, and due process. Governance must therefore include documentation of reward criteria, transparency about trade-offs, and mechanisms for appeal when automated planning affects rights or access to services.

Infrastructure is a necessary but often underappreciated dimension of governance. Reward-shaped planning depends on data pipelines, annotation platforms, model registries, evaluation harnesses, monitoring dashboards, incident response systems, and compute resources. Each component has operational and environmental costs. The foundation model literature notes that large-scale models concentrate capability and risk while requiring substantial

computational and organizational resources [19]. Sustainability concerns include energy consumption, water use for cooling, hardware supply chains, and the unequal distribution of environmental burdens. A reward system that optimizes for short-term task success may inadvertently encourage excessive model calls, redundant planning, or unnecessary data collection. Sustainable deployment requires lifecycle assessment and incentives for efficiency, reuse, and graceful degradation. It also requires institutions to plan for maintenance, retraining, and decommissioning, because reward models can become stale as environments and social norms change.

Accountability in reward-shaped systems is complicated by distributed agency. A plan may be generated by a language model, filtered by a reward model, constrained by a rule engine, and approved by a human operator. When harm occurs, responsibility can diffuse across the pipeline. The machine behaviour literature argues that artificial agents operate within social and institutional systems and can shape human behavior in return [25]. Governance frameworks should therefore assign clear responsibilities for reward specification, validation, monitoring, and redress. Auditing should include not only model outputs but also the reward pipeline, the data used to train evaluators, the selection of evaluation metrics, and the mechanisms by which affected parties can challenge decisions. Without such accountability, reward shaping risks becoming a technocratic substitute for democratic deliberation.

## 5. Robustness, Fairness, and Sustainability

Robustness in reward-shaped planning refers to the ability of a system to maintain acceptable behavior under uncertainty, distribution shift, adversarial manipulation, and operational failure. Reward hacking is a persistent concern: agents may discover shortcuts that achieve high reward without fulfilling the intended goal [1]. In language models, reward hacking can appear as sycophancy, verbose but unhelpful explanations, fabricated citations, or strategic omission of uncertainty. Specification gaming can also occur when a reward model overvalues surface features such as politeness, confidence, or formatting. Robustness requires adversarial testing, red-teaming, uncertainty quantification, and fallback procedures. It also requires reward models that are sensitive to context and capable of recognizing when a task is outside their competence. A system that confidently plans beyond its knowledge boundaries can be more dangerous than one that abstains or escalates to human review.

Fairness in reward-shaped planning involves distributive, procedural, and representational concerns. Distributive fairness asks whether benefits and burdens are allocated equitably across groups. Procedural fairness concerns whether affected parties have voice in how reward criteria are defined and applied. Representational fairness concerns whose perspectives are included in training data, annotation, and evaluation. Research on stochastic parrots and large language models warns that scale alone does not ensure equitable outcomes and can reproduce historical inequalities [23]. Responsible machine learning in healthcare further illustrates that fairness must be integrated into the full lifecycle, from problem formulation to post-deployment monitoring [24]. Reward shaping can support fairness when it explicitly rewards equitable outcomes and inclusive participation, but it can also undermine fairness when proxy metrics correlate with protected attributes or when optimization favors majority preferences. Fairness auditing must therefore examine reward models, training data, deployment contexts, and the distribution of error across groups.

Sustainability has both environmental and institutional dimensions. Environmentally, reward-shaped planning may require repeated sampling, large reward model inference, and extensive human annotation. These costs can be justified for high-value decisions, but they can also

become routine and wasteful. Institutionally, sustainability refers to the ability of organizations to maintain reward pipelines over time. Reward models require version control, documentation, retraining, and incident management. If these practices are neglected, systems may degrade silently or encode outdated norms. Sustainability also involves labor conditions for data annotators and domain experts, who may face psychological harm, low pay, and precarious employment. A sustainable reward ecosystem must account for these human costs and provide fair compensation, support, and career pathways. Without such considerations, reward shaping becomes a form of invisible infrastructure that externalizes risk and labor.

## 6. Case Illustrations and Cross-Domain Comparisons

Healthcare illustrates both the promise and peril of reward-shaped language planning. A language model might assist with triage, care coordination, or clinical documentation. Reward signals could prioritize accuracy, safety, equity, and adherence to clinical guidelines. However, healthcare decisions involve uncertainty, patient preferences, and legal liability. A reward model that overvalues efficiency might understate risks or overlook social determinants of health. A reward model that overvalues caution might generate excessive referrals or delay care. The responsible machine learning roadmap for healthcare emphasizes stakeholder engagement, fairness evaluation, and monitoring across deployment [24]. In this context, reward shaping should be constrained by clinical governance and patient rights. The system must be able to explain its recommendations, defer to clinicians, and document how reward criteria were selected and validated.

Public administration offers another instructive case. Governments may use language models to draft policy summaries, assess benefits eligibility, or coordinate emergency response. Reward shaping could reward transparency, legal compliance, and equitable service delivery. Yet public decisions must satisfy administrative law, due process, and democratic accountability. A reward model trained on historical decisions may reproduce discriminatory patterns or prioritize institutional convenience over citizen rights. The machine behaviour perspective is relevant because automated systems can reshape the behavior of public officials and citizens alike [25]. Governance should therefore include public consultation, freedom of information, and independent oversight. Reward criteria should be published in accessible form, and affected individuals should have meaningful opportunities to contest automated recommendations.

Scientific discovery and enterprise workflows provide contrasting cases. In scientific discovery, reward shaping may reward novelty, reproducibility, and empirical validation. Tool-using agents can search literature, generate hypotheses, run simulations, and analyze data [14]. Reflexive agents can revise plans based on experimental feedback [15]. However, scientific reward systems are vulnerable to publication bias, p-hacking, and overfitting to benchmarks. In enterprise settings, reward shaping may reward revenue, customer satisfaction, or operational efficiency. These objectives can conflict with privacy, labor rights, and long-term resilience. Cross-domain comparison suggests that reward shaping is always domain-specific and value-laden. There is no universal reward function for structured planning. Instead, there are domain-specific governance arrangements, evaluation cultures, and accountability mechanisms that must be integrated into system design.

## 7. Policy Implications and Forward-Looking Perspectives

Policy for reward-shaped language models should move beyond voluntary principles toward enforceable standards, audit requirements, and liability rules. Regulators may require

documentation of reward criteria, training data provenance, evaluation results, and known limitations. They may also require human oversight for high-risk decisions and provide rights to explanation and appeal. International coordination is necessary because language models operate across borders, but local values and legal traditions must be respected. A one-size-fits-all reward framework is unlikely to be legitimate or effective. Instead, policy should support layered governance: technical standards for safety and robustness, sectoral rules for healthcare and finance, and democratic oversight for public-sector use. Research on scalable agent alignment suggests that reward modeling can be part of a broader alignment strategy, but it cannot substitute for institutional accountability [17].

Future systems may incorporate more sophisticated reward shaping through multi-agent debate, formal verification, causal modeling, and participatory design. However, greater sophistication does not automatically produce greater legitimacy. A technically advanced reward model may still encode narrow interests if affected communities are excluded from its design. Policy should therefore invest in public capacity to evaluate and contest reward-shaped systems. This includes funding independent audits, supporting civil society participation, and developing educational programs for domain experts. It also includes protecting whistleblowers and researchers who identify reward hacking or discriminatory outcomes. The goal is not to eliminate reward shaping but to make its value choices visible, contestable, and revisable. In this sense, reward shaping for structured planning is inseparable from the politics of automation and the distribution of decision-making power.

## 8. Conclusion

Reward shaping for structured planning and decision-making in large language models is a systems-level challenge that spans algorithm design, infrastructure, governance, and social legitimacy. Reward signals guide how language models decompose tasks, select tools, allocate attention, and revise plans. They can improve long-horizon coherence, alignment with human preferences, and robustness in complex environments. They can also introduce reward hacking, specification gaming, opacity, bias, and unsustainable resource consumption. The analysis in this paper has argued that these outcomes are not determined by reward functions alone. They emerge from architectural choices, data pipelines, evaluation cultures, organizational incentives, and regulatory contexts. A centralized reward model may simplify governance but concentrate power. A decentralized system may improve contextual fit but complicate accountability. Process rewards may improve reasoning quality but increase annotation burdens. Outcome rewards may be easier to verify but vulnerable to lucky guesses and proxy failure.

The paper has also emphasized that fairness and sustainability are not optional additions to reward-shaped planning. They are constitutive of whether such systems deserve trust. Fairness requires attention to distributive, procedural, and representational concerns across the lifecycle. Sustainability requires attention to environmental costs, labor conditions, maintenance obligations, and institutional resilience. Governance requires documentation, auditability, contestability, and clear responsibility. Policy should support layered oversight that combines technical standards, sectoral regulation, and democratic participation. Future research should develop methods for multi-objective reward design, robust evaluation under distribution shift, participatory reward specification, and lifecycle impact assessment. It should also examine how reward-shaped language agents interact with human decision-makers and institutions over time. Ultimately, reward shaping for structured planning should be treated as a socio-technical design problem. Its success will depend not only on better

optimization but also on whether its value choices are transparent, inclusive, and accountable to those affected by automated decisions.

**References**


1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307). Curran Associates.

3. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

4. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

5. Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In Proceedings of the Sixteenth International Conference on Machine Learning (pp. 278-287). Morgan Kaufmann.

6. Russell, S. J., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008). Curran Associates.

8. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (pp. 1877-1901). Curran Associates.

9. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (pp. 27730-27744). Curran Associates.

10. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., ... Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

11. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (pp. 24824-24837). Curran Associates.

13. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In International Conference on Learning Representations. OpenReview.

14. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. In Advances in Neural Information Processing Systems (pp. 68539-68551). Curran Associates.

15. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In Advances in Neural Information Processing Systems (pp. 8634-8652). Curran Associates.

16. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In 2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE) (pp. 29-32). IEEE.

17. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. arXiv preprint arXiv:1811.07871.

18. Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. In International Conference on Learning Representations. OpenReview.

19. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

20. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

21. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (pp. 2503-2511). Curran Associates.

22. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15). ACM.

23. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM.

24. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Snoek, J., Schulam, P., Rajkomar, A., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.

25. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., ... Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477-486.