

Curriculum Reinforcement Learning for Progressive Complex Reasoning in Large Language Models

Yashvi Tank

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
yashvi1989@buffalo.edu

Scarlett R. Johnstone

Department of Computer Science, George Mason University, Fairfax, VA, USA.
johnstone1987@gmu.edu

Malwina Wiaderek

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
mwiaderek@ku.edu

Pascale Rudolph

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
pascalerudolph87@binghamton.edu

Abstract

Large language models have demonstrated remarkable capacity for pattern completion, instruction following, and multi-step reasoning, yet their ability to solve progressively complex tasks remains uneven, brittle, and dependent on hidden interactions among data, reward design, compute allocation, and deployment context. This paper examines curriculum reinforcement learning as a system-level strategy for cultivating progressive complex reasoning in large language models. It argues that reasoning capability is not merely an emergent property of parameter scale but an outcome of orchestrated task sequencing, feedback governance, policy optimization, verification, observability, and socio-technical safeguards. The discussion integrates curriculum learning, reinforcement learning from human feedback, chain-of-thought prompting, preference optimization, and scalable policy improvement into a layered architecture. It analyzes structural trade-offs among sample efficiency, training stability, interpretability, cost, latency, fairness, and sustainability. The paper then considers infrastructure requirements, data provenance, annotation labor, synthetic data risk, red-teaming, monitoring, incident response, and policy implications. Case illustrations from education, scientific discovery, public-sector decision support, and enterprise assistance show that staged reasoning curricula can improve robustness and transfer, but they can also amplify dependency on reward models, obscure accountability, and shift labor burdens. The conclusion calls for auditable curricula, participatory governance, lifecycle evaluation, and sustainability-aware deployment as foundational requirements for trustworthy progressive reasoning systems.

Keywords

curriculum reinforcement learning; large language models; complex reasoning; AI governance; socio-technical systems; robustness; fairness; sustainable AI infrastructure.

1. Introduction

The modern generation of large language models rests on transformer architectures and large-scale pretraining [1], [2]. These systems acquire broad linguistic and factual regularities from heterogeneous corpora, enabling few-shot adaptation, instruction following, and interactive problem solving. Chain-of-thought prompting and related elicitation techniques have further shown that model performance on reasoning tasks can improve when intermediate reasoning steps are generated or encouraged [3], [4]. Reinforcement learning from human feedback and instruction tuning have become central mechanisms for aligning model behavior with user expectations, safety constraints, and task-specific goals [5], [6]. Nevertheless, the path from fluent generation to reliable complex reasoning remains contested. Models can solve benchmark problems while failing on modest perturbations, produce plausible but invalid rationales, and exhibit uneven performance across domains, languages, and populations. These failures are not purely algorithmic; they emerge from system design, data curation, evaluation regimes, infrastructure choices, and governance arrangements.

Curriculum reinforcement learning offers a promising but demanding response to this problem. Instead of exposing a model to the full distribution of tasks at once, curriculum reinforcement learning sequences experiences according to difficulty, prerequisite structure, feedback availability, or competence growth. The approach draws on earlier work in curriculum learning and staged neural network development, which demonstrated that ordering examples can affect convergence, generalization, and stability [8], [9]. In the context of large language models, curriculum reinforcement learning can be understood as a socio-technical control layer that coordinates task generation, reward modeling, policy optimization, verification, and deployment feedback. It is not simply a training trick. It is an architecture for managing complexity under uncertainty. Recent work on scalable reasoning path filtering and alignment for small-parameter large language models illustrates the growing interest in making staged policy improvement more efficient and controllable [7]. That work is relevant not because it solves all reasoning problems, but because it highlights a broader systems question: how can organizations design reinforcement learning pipelines that progressively build reasoning competence without sacrificing transparency, fairness, or operational sustainability?

This paper develops that question across conceptual, architectural, infrastructural, evaluative, and policy dimensions. It first reviews the foundations of curriculum learning and reinforcement learning for language models. It then proposes a layered architecture for progressive complex reasoning, emphasizing the interaction among task generators, curriculum controllers, reward models, verifiers, memory systems, and observability tools. It examines training infrastructure and data governance, including annotation labor, provenance, privacy, and synthetic data risk. It analyzes robustness, fairness, and evaluation as coupled concerns rather than separate checklist items. It considers deployment, sustainability, and policy implications, especially the costs of continuous reinforcement learning and the governance of models that make consequential recommendations. It then presents cross-domain case illustrations and a forward-looking research agenda. The central claim is that progressive complex reasoning requires not only better optimization but also institutional design that makes curricula auditable, feedback contestable, and deployment accountable.

2. Conceptual Foundations of Curriculum Reinforcement Learning

Curriculum learning is grounded in the intuition that learning is easier when the order of experience respects developmental or structural constraints. Early work showed that neural networks can benefit from starting with simpler examples before moving to harder ones,

particularly when the target function has hierarchical or compositional structure [9]. Later formulations generalized this idea into a family of strategies for ordering, weighting, or selecting training examples according to difficulty, diversity, or learning progress [8]. In language modeling, curriculum principles appear in pretraining data mixing, instruction tuning stages, chain-of-thought fine-tuning, and reinforcement learning from human feedback. The appeal is straightforward: if a model receives tasks that are too difficult too early, it may learn spurious shortcuts, suffer unstable gradients, or overfit to surface patterns. If it remains too long in easy regions, it may fail to develop robust abstraction and compositional generalization.

Reinforcement learning adds a different set of concerns. In reinforcement learning, an agent improves by interacting with an environment and receiving rewards. Classic successes in game playing and control demonstrate the power of trial-and-error learning when environments are well specified and feedback is reliable [11]. Proximal policy optimization and related methods provide practical mechanisms for stabilizing updates by constraining how far a policy changes in a single iteration [12]. Neural architecture search with reinforcement learning further shows how search and optimization can be combined when the design space is large and discrete [10]. When these ideas are transferred to large language models, however, the environment is not a game with a clear score. It is a mixture of human preferences, task rubrics, verifiers, retrieval systems, simulators, and deployment contexts. Rewards are often learned from human comparisons, which introduces noise, bias, and strategic ambiguity [6]. The result is a hybrid system in which curriculum, reward modeling, and policy optimization are entangled.

Reinforcement learning from human feedback has become a dominant alignment paradigm because it allows developers to shape model behavior using preference data rather than hand-coded rules [5], [6]. Related work on helpful and harmless assistants shows that reinforcement learning can incorporate safety constraints and refusal behavior alongside task performance [13]. Alignment research has also attempted to formalize value pluralism and moral uncertainty, recognizing that human preferences are not monolithic [14]. These contributions matter for curriculum reinforcement learning because they imply that difficulty is not the only dimension along which tasks should be sequenced. Safety, ambiguity, cultural context, uncertainty, and contestability also shape whether a reasoning task is appropriate at a given stage. A curriculum that advances only on accuracy may produce a model that is increasingly capable but less governable. A curriculum that advances on capability, calibration, refusal, and explanation quality may produce slower but more trustworthy progress.

The conceptual foundation therefore requires a shift from single-objective training to multi-dimensional progression. In practice, curriculum reinforcement learning for large language models must decide what counts as simple, what counts as complex, and who has authority to make that judgment. Arithmetic word problems may be simple in one sense but culturally specific in another. Legal reasoning may be complex, but its complexity is not merely length or syntax; it involves precedent, jurisdiction, ambiguity, and institutional consequence. Scientific reasoning involves hypothesis generation, experimental design, and causal inference. Public-sector decision support involves contested values and legal accountability. A robust curriculum must represent these dimensions explicitly, even if the operational implementation remains approximate. Otherwise, the system may optimize for benchmark difficulty while ignoring the social and epistemic structures that make complex reasoning meaningful.

3. Architecture and System Design for Progressive Reasoning

A system-level architecture for curriculum reinforcement learning should separate concerns while maintaining feedback across layers. At the base is a model serving and training substrate that supports pretraining checkpoints, instruction-tuned variants, policy optimization, and inference. Above that lies a task generation layer that produces prompts, environments, rubrics, and verification signals. A curriculum controller then sequences tasks according to learner state, domain prerequisites, uncertainty estimates, and governance constraints. A reward and preference layer converts human feedback, model-based judgments, verifier outcomes, and safety criteria into training signals. A verification layer checks reasoning traces against external tools, knowledge bases, symbolic solvers, or human review. A memory and retrieval layer provides context, provenance, and state across long-horizon tasks. Finally, an observability and audit layer records decisions, data lineage, reward distributions, failure modes, and intervention points. This layered view helps avoid the common mistake of treating curriculum reinforcement learning as a single training loop.

The curriculum controller is the most distinctive component. It must balance exploration and exploitation at the level of tasks rather than tokens. Early stages may emphasize synthetic, well-verified tasks with short reasoning chains and deterministic rewards. Middle stages may introduce multi-step problems, retrieval, tool use, and partial feedback. Later stages may involve open-ended tasks, adversarial scenarios, and value-laden dilemmas. The controller can use signals such as success rate, calibration, explanation fidelity, refusal accuracy, and compute cost to decide when to advance. However, these signals are incomplete. A model may succeed for the wrong reasons, refuse appropriately for the wrong reasons, or explain a correct answer with invalid reasoning. Verification is therefore essential. Without verification, curriculum progression can become a performance ritual that rewards superficial compliance.

Reward design is equally consequential. Human preference data are expensive and noisy, and annotators may disagree because tasks are ambiguous or values conflict [14]. Learned reward models can be gamed, especially when the policy discovers regions of high reward that are not aligned with human intent [18]. Curriculum reinforcement learning can mitigate this risk by introducing reward diversity gradually, but it can also amplify it if early reward models encode narrow cultural assumptions. A mature architecture should include multiple reward sources: human preferences, expert rubrics, verifier outcomes, safety classifiers, and deployment telemetry. These sources should not be collapsed too early into a single scalar. Maintaining separable signals allows auditors to inspect which objectives drive behavior and to reweight them when new evidence emerges.

Recent work on scalable reasoning path filtering and alignment for small-parameter models points toward a pragmatic middle ground [7]. Rather than relying only on massive models and monolithic reward models, such approaches filter reasoning trajectories, align them selectively, and focus compute on promising paths. In system terms, this suggests that curriculum reinforcement learning can benefit from path-level curation, not just prompt-level sequencing. The architecture should therefore treat reasoning traces as first-class artifacts with metadata, verification status, and provenance. This creates opportunities for reuse, audit, and targeted improvement. It also introduces risks: filtered traces may overrepresent easy-to-verify reasoning styles, exclude culturally specific argumentation, or reinforce the preferences of the filtering model. Governance must address these risks through sampling audits, diversity constraints, and human review of excluded cases.

4. Training Infrastructure and Data Governance

The infrastructure required for curriculum reinforcement learning is substantial. It includes distributed training clusters, high-throughput inference for rollouts, retrieval services, tool execution sandboxes, simulation environments, data versioning systems, experiment tracking, and monitoring. Unlike standard supervised fine-tuning, reinforcement learning pipelines generate many candidate outputs and require repeated evaluation. This increases compute intensity, storage needs, and operational complexity. The hidden technical debt of machine learning systems becomes especially relevant because curriculum reinforcement learning introduces tight coupling among data pipelines, reward models, policy checkpoints, and evaluation code [19]. A change in the reward model can invalidate prior curriculum decisions. A shift in the task generator can alter difficulty distributions. A new safety classifier can block trajectories that were previously used for training. Without strong versioning and dependency management, reproducibility deteriorates quickly.

Data governance is therefore not a compliance afterthought but a core architectural requirement. Training corpora, synthetic tasks, human annotations, preference comparisons, and verification logs all carry provenance and licensing obligations. Foundation models already raise concerns about data concentration, labor, and environmental cost [15], [16]. Curriculum reinforcement learning intensifies these concerns because it may require domain experts, red-teamers, and affected communities to produce and evaluate tasks over time. Annotation labor must be compensated, documented, and protected. Privacy must be preserved when tasks involve personal, medical, legal, or financial information. Consent and purpose limitation must be enforced across data reuse. Synthetic data can alleviate scarcity, but it can also create feedback loops in which model-generated errors become training targets. Governance should therefore include synthetic data audits, contamination checks, and clear boundaries between simulated and verified knowledge.

A further governance challenge is the distribution of authority. Who decides that a model is ready to advance from one curriculum stage to another? In commercial settings, product teams may prioritize engagement or cost reduction. In public-sector settings, agencies may prioritize legal defensibility and equity. In scientific settings, domain experts may prioritize validity and reproducibility. Curriculum reinforcement learning makes these choices explicit because advancement criteria must be encoded. If the criteria are hidden in code or reward weights, accountability becomes difficult. If they are published, they may be gamed or politicized. A defensible approach is to maintain an auditable curriculum specification that records task categories, difficulty metrics, advancement thresholds, override procedures, and appeal mechanisms. This specification should be reviewed by interdisciplinary committees that include technical, legal, ethical, and domain expertise.

5. Robustness, Fairness, and Evaluation

Robustness in progressive complex reasoning means more than high accuracy on a static benchmark. It means maintaining performance under distribution shift, adversarial prompting, ambiguous instructions, missing context, tool failure, and value conflict. Chain-of-thought and zero-shot reasoning methods have improved benchmark performance, but they also reveal sensitivity to prompt phrasing, exemplar choice, and evaluation conditions [3], [4]. Curriculum reinforcement learning can improve robustness by training models on controlled variations, gradually introducing noise and ambiguity, and requiring verification before advancement. However, robustness gains can be superficial if the curriculum overfits to the same evaluators used for advancement. Independent red-teaming, held-out task families, and

longitudinal evaluation are necessary. Robustness should be measured across time, domains, languages, and user populations.

Fairness is intertwined with robustness because models that fail unevenly can produce discriminatory outcomes. Foundation models can encode historical biases, marginalize dialects, and underperform for underrepresented groups [15], [16]. In curriculum reinforcement learning, fairness risks arise at several points. Task generation may overrepresent dominant cultural contexts. Reward models may penalize unfamiliar reasoning styles. Verification tools may lack coverage for minority languages or informal knowledge systems. Advancement criteria may optimize for average performance while ignoring variance across groups. Alignment frameworks that emphasize shared human values can inadvertently flatten moral disagreement [14]. A fair curriculum must therefore include disaggregated evaluation, participatory task design, and mechanisms for contesting model outputs. Fairness is not achieved by adding a single constraint; it requires ongoing governance of the curriculum's boundary conditions.

Evaluation must also address explanation quality and calibration. A model may produce a correct answer with an unfaithful rationale, or an incorrect answer with a confident rationale. Curriculum reinforcement learning can reward explanation fidelity by requiring traces to pass verifier checks or expert review. It can reward calibration by penalizing overconfidence when evidence is weak. It can reward refusal when tasks exceed competence or violate policy. These dimensions should be tracked alongside accuracy because they determine whether a system is trustworthy in deployment. Evaluation should include adversarial cases, counterfactual tests, and scenario-based assessments that reflect real-world consequences. The goal is not to prove that a model reasons like a human, but to establish whether its behavior is reliable enough for a specified context and whether its limitations are visible to users and overseers.

6. Deployment, Sustainability, and Policy Implications

Deployment changes the meaning of curriculum reinforcement learning. In a laboratory, a curriculum is a training strategy. In deployment, it becomes a lifecycle process. Models encounter new tasks, user populations, regulations, and adversarial pressures. Continuous learning may be desirable, but it introduces risks of drift, reward hacking, and unauthorized capability expansion. Deployment architectures should therefore separate stable inference models from experimental training pipelines, with strict controls on what feedback can influence production behavior. Monitoring should track performance, refusal rates, latency, cost, and disparate impact. Incident response should allow rapid rollback, investigation, and disclosure. Governance should define who can approve curriculum changes, how affected users are notified, and how appeals are handled.

Sustainability is a first-order concern. Large language models already consume significant energy and water for training and inference, and reinforcement learning pipelines multiply these costs through repeated rollouts and evaluations [17]. Curriculum reinforcement learning can improve sample efficiency by focusing compute on appropriate tasks, but it can also increase total compute if advancement criteria are poorly designed or if multiple reward models are run simultaneously. Sustainable design requires measuring energy use, carbon intensity, hardware utilization, and marginal value per unit of compute. It also requires questioning whether every reasoning capability needs to be trained continuously. Periodic retraining, distillation, caching, retrieval augmentation, and smaller specialized models may be more sustainable than endlessly scaling monolithic reinforcement learning. Recent work on

small-parameter alignment and reasoning path filtering suggests that efficiency-oriented architectures deserve serious attention [7].

Policy implications extend beyond environmental impact. Governments and standards bodies are increasingly interested in auditing, transparency, and risk management for foundation models [16]. Curriculum reinforcement learning complicates these efforts because model behavior emerges from a sequence of training decisions that may be distributed across organizations. A regulator may inspect a final model but lack access to the reward models, task generators, annotation protocols, and advancement logs that shaped it. Policy should therefore require documentation of curricula for high-risk systems, including task categories, reward sources, evaluation results, and known limitations. It should also protect whistleblowers and affected communities who report harms. International coordination is needed because models and data cross borders, but local values and legal traditions must be respected. The policy goal is not to freeze innovation but to make progressive reasoning systems legible, contestable, and accountable.

7. Case Illustrations and Cross-Domain Comparisons

In education, curriculum reinforcement learning could support tutoring systems that gradually introduce problem types, diagnose misconceptions, and adapt explanations. The appeal is strong because education already uses scaffolding and prerequisite structures. However, the risks are equally strong. A model may optimize for answer completion rather than learning, reward compliance rather than curiosity, or impose a single pedagogical style. Fairness concerns arise when curricula reflect dominant linguistic or cultural norms. Governance should involve teachers, students, and parents in task design and evaluation. The system should preserve teacher oversight and avoid displacing professional judgment. A successful educational curriculum would measure learning gains, calibration, motivation, and equity, not just task accuracy.

In scientific discovery, curriculum reinforcement learning could help models progress from literature synthesis to hypothesis generation, experimental design, and interpretation. Scientific reasoning is verifiable in some domains through simulation, statistical analysis, and replication, but it is also creative and contested. A curriculum could start with well-structured data analysis and move toward open-ended discovery. Yet scientific validity depends on domain expertise, instrumentation, and peer review. A model that advances through internally generated verification may develop confident but unsupported claims. Infrastructure must integrate lab information systems, provenance tracking, and expert review. Cross-domain comparison shows that education and science share a need for staged competence, but they differ in verification regimes and accountability structures.

In public-sector decision support, curriculum reinforcement learning could assist with benefits eligibility, zoning analysis, environmental review, or emergency response. These domains involve legal rules, contested values, and high-stakes consequences. A curriculum might begin with information retrieval and procedural guidance, then advance to scenario analysis and recommendation drafting. But the model should not make final decisions. Governance should require human decision-makers, audit trails, and appeal rights. Fairness is central because errors can deny benefits or exacerbate inequality. Robustness is also central because policy changes and emergencies create distribution shifts. Here, the curriculum must be tightly coupled to legal review and public accountability, not merely to benchmark performance.

In enterprise assistance, curriculum reinforcement learning can improve customer support, software engineering, and business analytics. These settings offer abundant feedback and relatively clear metrics, making them attractive for staged training. However, they also create incentives for engagement optimization, labor displacement, and hidden surveillance. A curriculum that rewards fast resolution may penalize careful escalation. A curriculum that rewards user satisfaction may ignore long-term trust. Cross-domain comparison suggests that the same technical architecture can produce very different outcomes depending on institutional context. Therefore, curriculum reinforcement learning should be evaluated as a governance intervention, not only as a modeling technique.

8. Future Directions and Research Agenda

Several research directions follow from this analysis. First, curriculum specification languages are needed to make task categories, difficulty dimensions, reward sources, and advancement criteria explicit and auditable. These specifications should be machine-readable for operations but human-interpretable for oversight. Second, verification architectures should combine symbolic tools, retrieval, simulation, and expert review in ways that are robust to domain shift. Third, reward modeling should move beyond single scalar preferences toward multi-objective, contestable representations that preserve disagreement and uncertainty [14], [18]. Fourth, evaluation should be longitudinal and participatory, involving affected communities and independent auditors. Fifth, infrastructure should be designed for sustainability, with metrics for energy, carbon, water, and hardware efficiency integrated into curriculum decisions [17].

A sixth direction concerns multi-agent and human-in-the-loop systems. Progressive complex reasoning may be better supported by ensembles of specialized models, tools, and human experts than by a single monolithic policy. Curriculum reinforcement learning could coordinate these components, assigning tasks according to competence and confidence. However, multi-agent systems introduce new failure modes, including collusion, error propagation, and accountability diffusion. Governance must specify responsibility across components. A seventh direction is the development of smaller, efficient models that can be aligned through filtered reasoning paths and staged reinforcement learning [7]. This may reduce cost and increase deployability, but it also requires careful evaluation of capability limits. Finally, policy research should examine how curriculum documentation can be standardized without creating burdens that favor large firms. Open standards, independent audits, and public registries may help balance transparency and innovation.

9. Conclusion

Curriculum reinforcement learning for progressive complex reasoning in large language models is best understood as a socio-technical architecture rather than a single training algorithm. It brings together task sequencing, reward modeling, policy optimization, verification, observability, data governance, deployment controls, and policy oversight. The promise is significant: by staging difficulty, diversifying feedback, and verifying progress, systems may become more sample-efficient, robust, and trustworthy. The risks are equally significant: hidden curricula can encode narrow values, reward models can be gamed, annotation labor can be exploited, compute costs can escalate, and accountability can dissolve across distributed pipelines. The required reference at position seven illustrates one technical direction toward scalable filtering and alignment, but the broader lesson is that efficiency gains must be embedded in governance. The field needs auditable curricula, disaggregated evaluation, participatory design, sustainability metrics, and clear legal responsibility. Without

these, progressive reasoning systems may become more capable without becoming more legitimate. With them, curriculum reinforcement learning can contribute to language models that reason more carefully, fail more transparently, and serve a broader range of human purposes.

References

1. 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* (Vol. 30, pp. 5998–6008). Curran Associates.
2. 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* (Vol. 33, pp. 1877–1901).
3. 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* (Vol. 35, pp. 24824–24837).
4. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 22199–22213).
5. 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* (Vol. 35, pp. 27730–27744).
6. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4299–4307).
7. 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.
8. Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 41–48). ACM.
9. Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1), 71–99.
10. Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*.
11. 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.

12. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
13. 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, T., ... Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
14. 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.
15. 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.
16. 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.
17. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). ACL.
18. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
19. 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 (Vol. 28, pp. 2503–2511).
20. Whittlestone, J., Nyrupe, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 195–200). ACM.