

~~Self-Improving AI~~AI & Human Co-Improvement
for Safer Co-Superintelligence

Jason Weston, Jakob Foerster

FAIR at Meta

Self-improvement is a goal currently exciting the field of AI, but is fraught with danger, and may take time to fully achieve. We advocate that a more achievable and better goal for humanity is to maximize co-improvement: collaboration between human researchers and AIs to achieve co-superintelligence. That is, specifically targeting improving AI systems’ ability to work with human researchers to conduct AI research together, from ideation to experimentation, in order to both accelerate AI research and to generally endow both AIs and humans with safer superintelligence through their symbiosis. Focusing on including human research improvement in the loop will both get us there faster, and more safely.

~~The quest for Self-Improving AI~~

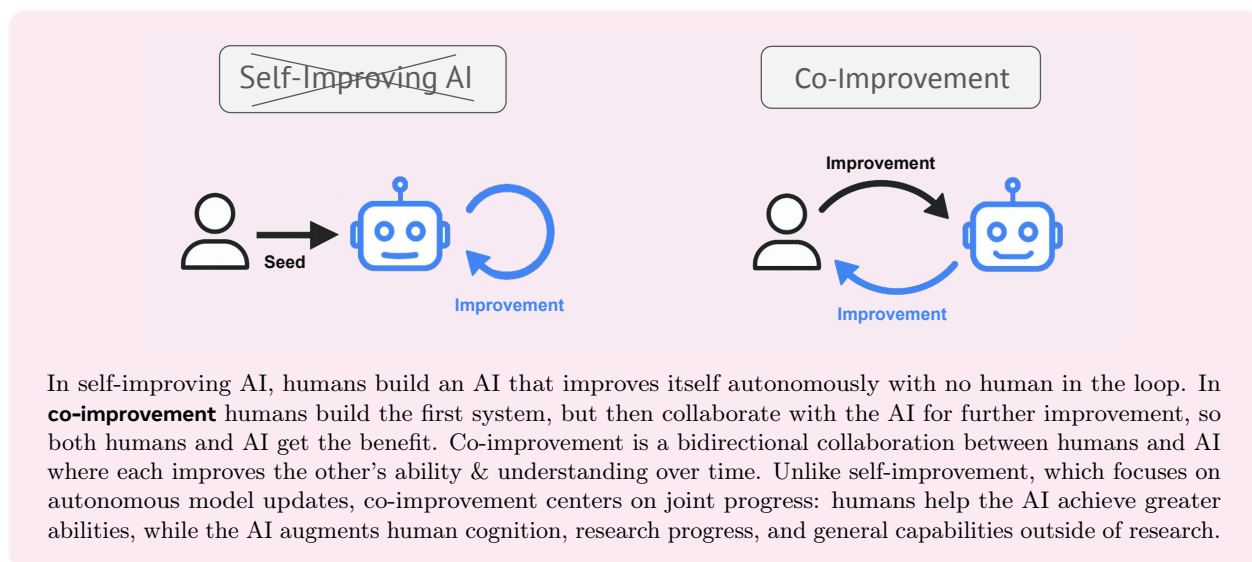
AI that improves itself has been the primary goal of the field since its inception [1]. Historically, practical instantiations have focused on parameterizing a model in terms of weights, and then finding the best choice of those weights, without human intervention – from linear models through to neural networks. The 2010s onwards marked the era of scaling to ever-larger models [2], yielding large performance gains, but still focused on self-improving via weights only, while the architecture, data, objective function, update rule, and implementation (code) were mostly fixed. The current era has expanded that self-improvement search to improving all aspects through learning: models that create their own training data [3], challenge themselves to be better [4, 5], and learn to evaluate and reward themselves based on their own performance on these tasks [6, 7, 8]; see Table 2. Some of these axes have already provided sizable gains in performance, and approaches like synthetic data creation and LLM-as-a-Judge are now standard building blocks for frontier models. The quest for AI that improves its own architectures and rewrites its own code is still in its infancy [9, 10], but early signs show its promise, with a current push on autonomous AI research agents [11, 12, 13, 14].

It seems clear by now that we are marching towards ever more intelligent AI systems that in the long run will surpass humans in all task metrics, and by a large margin. Fully realized self-improvement is clearly an end-game marker. However, endowing AIs with this autonomous ability without appropriate guidance built into the system is fraught with danger for humankind – from misuse through to misalignment [15]. Nevertheless, there is still time left before AI eclipses humans in all endeavors, and in AI research in particular. We thus propose that it is better to focus on humans and AI working together to solve these problems.

A better quest for humans: Co-Improving AI

Our central position is that **“Solving AI” is accelerated by building AI that collaborates with humans to solve AI**. This is distinct from the goal of self-improving AI, which seeks to eliminate humans from the loop as quickly as possible, where AI performs its own research and learning autonomously. Instead, we advocate for **co-improvement**, whereby collaborative AI agents are built with the goal of conducting research *with* humans. Thus, we accelerate the research *with the research*. Importantly, including humans in the loop allows us the ability to steer the research in the right directions, i.e. “Solving AI” means a *positive solution for humanity*. In particular, we believe that such a *positive solution* is one where AI augments and enables humans in all areas of society, rather than pursuing full automation that removes human decision-making.

While today most of the AI research is done by humans, we expect over time the burden will become more shared – as AI improves it can work with us to take more of the workload in providing solutions. Collaboration can take advantage of the complementary skill sets of humans and AI, which currently excel in quite different



areas, while we expect over time AI will continue to outpace us in more and more dimensions. However, because AI is not mature enough yet to fully self-improve and is susceptible to misalignment, we expect that co-improvement will get us there faster *and* more safely. That is, with the help of AI we are more likely to solve the capability and safety problems of AI — but with humans in the loop, collaborating on the research. Thus, co-improvement can help induce positive outcomes for humanity.

Aren't we already doing this? There are multiple plausible pathways to superintelligence, and current AI systems are already helping us explore these research directions to some degree, e.g. through code assistance and writing assistance. And, in general, improving the overall capabilities of frontier models does equip them with some of the skills suitable for research collaboration – as a byproduct. However, we typically see that skills improve further when they are targeted, for example lots of effort has been spent on improving AI coding skills, resulting in improved AI coding ability. However, while coding isn't solved either [16], there is much more to “solving” AI than coding alone. Our central point is that with more development resources spent on endowing AI with AI research collaboration skills, these abilities will get better too.

What can we gain? Progress in AI has been made with a combination of both training data and method changes from architecture through to training objectives, often with these advances working in tandem, leading to notable paradigm shifts. For example, the creation of Imagenet and the introduction of AlexNet [33, 34], curating web data and scaling transformers [35, 36, 37], the labeling of instruction following data and building of RLHF training [38, 39, 40], or the collection of verifiable reasoning tasks and the use of RLVR for training chain-of-thought [41, 42, 43, 25]. In each case it took human researchers significant effort, with many smaller intermediate results as well as wrong directions and dead ends, in order to find these wins. Any improvement in our ability to do research will speed up this process. Hence, co-research with strong AI systems built to collaborate with us should accelerate finding the unknown new paradigm shifts which are currently missing.

Overall, we expect co-improvement can provide: (i) faster progress to find important paradigm shifts; (ii) more transparency and steerability than direct self-improvement in making this progress; (iii) more focus on human-centered safe AI. For example, we may be able to develop systems that are super-human at ML theory, so we could have provably safe AI. In contrast, an entirely autonomous AI self-improvement system can suffer from goal misspecification (e.g. what it means to “solve AI” does not take human needs into account).

How do we do it? In order to build AI that can collaborate with us on research, we should put some of our focus on building AI possessing these skills. So, that means measuring the research collaboration skills of AI with new benchmarks, and constructing training data and methods that improve these benchmarks, much as we do with building other skills. These skills should cover all major AI research activities that comprise the

Table 1 Co-improvement goals across major AI development activities, in order to achieve co-superintelligence.

Category	Mechanism	Category	Mechanism
Collaborative problem identification	Humans and AI help jointly define goals; identify current failures, brainstorm, propose unexplored directions, take into account existing work.	Safety & alignment	Humans and AI co-develop methods as well as values, and constitutions. Use whole research cycle listed (left) to develop and test them.
Benchmark creation & problem evaluation	Given identified problem, jointly define desiderata; benchmark construction & performance analysis; refine benchmarks to validate problem.	Systems & infrastructure co-design	Jointly architecting pipelines, optimizations, configs, and reproducibility improvements.
Method innovation & idea generation	Jointly brainstorm and identify solutions: systems, architectures, algorithms, training data, recipes and overall code designs for future models.	Integrating into real-world systems	Collaborating to convert research to real-world use; in turn this may identify further required research.
Joint experiment design	Co-design overall plan to test innovations: experiment protocol and setting, further benchmark identification, proposed ablations, ...	Scientific communication	Jointly drafting documentation & writeups, including figures and results; ensuring clarity of message and correctness.
Collaborative execution	Humans and AI co-produce and run multi-step workflows (implementation, experiments).	Collective intelligence & group research	Multi-human & AI collaboration for given problem; aggregate viewpoints, structure debate, and synthesize consensus & actionable steps.
Evaluation & error analysis	Analyzing performance both on benchmarks and individual cases for successes & failures (at scale via AI); feedback loop for research iteration.	Bidirectional co-improvement	Overall collaboration aims to enable increased intelligence in both humans & AI, including all manifested learnings from the research cycle, with the goal of achieving co-superintelligence.

Table 2 Major interpretations of **self-improvement** in AI research together with outstanding research challenges.

Category	Learnable axis					Mechanism	Representative Examples	Open Issues / Research Directions
	Parameters	Train data	Objective	Architecture	Overall code			
Classic parameter optimization	●	○	○	○	○	Model updates its own weights given data & fixed objective, i.e., classic training algorithms.	Gradient descent, SFT, LLM pre-training.	Data inefficiency; out-of-distribution generalization; compute inefficiency (model size, data size).
Self-training on own generations or actions	●	●	○	○	○	Model learns from its own generated outputs and reasoning traces; could use tools/actions in an environment.	RL with fixed/verifiable rewards [17, 18, 19, 20, 21, 22, 23, 24, 25]; STaR [26].	Computational efficiency; reward hacking; can drift away from human reasoning without strong priors.
Self-challenging / Self-play & Synthetic data creation	●	●	○	○	○	Model improves via self-generated tasks; can simultaneously improve its ability to generate tasks.	(CoT)-Self-Instruct [3, 27], Self-Challenging Agents [4], Absolute Zero [5], SPICE [28].	Task quality & correctness, diversity & generalization beyond synthetic tasks.
Self-evaluation / Self-reward	●	●	●	○	○	Model generates and applies its own feedback or reward signals; can learn reward model.	Self-Rewarding [7], Self-refinement, RLAI [29], Constitutional AI [6].	Ensuring value alignment; validating reliability of self-generated judgments.
Algorithm or architecture self-modification	●	○	●	●	●	Model changes its own structure, code, compiler transforms, or learning algorithm.	Neural arch. search [30]; code optimization, AlphaEvolve [31]; “AI scientist” [11, 12, 14].	Efficiency; ensuring safety and correctness; interpretability of modifications.
Recursive self-improvement	●	●	●	●	●	System designs successively more capable versions of itself, potentially achieving AGI/SI.	Gödel Machine [9]; theoretical AGI discussions [32].	Executing on this currently theoretical vision; ensuring safety.
Metaphorical or rhetorical usage	?	?	?	?	?	Non-technical use implying progressive capability or autonomy.	Media or industry narratives.	Clarifying terminology; avoiding conceptual confusion.

end-to-end research pipeline. We define some major ones in Table 1. These include collaboration with us to identify research problems, create training data and benchmarks, innovate methods, design and execute experiments, and conduct evaluation and error analysis which is then fed back to refine the whole process. Similarly, co-design and co-development of safety & alignment, systems improvements, integration of these innovations into real-world use cases, and scientific communication are also important goals. Crucially, in contrast to recent end-to-end AI scientist approaches [11, 14] the goal is to improve research quality, rather than to accelerate research artifact production (such as papers) via full automation.

From co-improvement to co-superintelligence We envision that the first goal of co-improvement is to improve our ability to conduct research into improving AI. We expect the end result when successful, like in the self-improving paradigm, will be a superintelligence system with ability to self-improve. However, the difference is when humans are working with the AI system to help achieve this at each step of the loop, we have more opportunities to guide this process towards positive benefits to humanity. In particular, we can consider safety and societal harms (see later section), and increase the collective knowledge of humanity at each step.

Going forward, we further envision that the goals of co-improvement could shift from building AI that collaborates on AI research toward co-improvement on *all kinds* of research or important topics to humanity. As AI becomes ever more capable, these new skills hopefully become even easier to attain. From the human societal standpoint, building AI can help *humans improve themselves, their abilities and knowledge, and their situation*. We can thus focus on building AI towards those goals. We thus refer to AI helping us achieve these abilities, beyond our current ones, as **co-superintelligence**, emphasizing what AI can give back to humanity.

Co-improvement and societal harms & benefits As capabilities increase the potential for harms can increase. Today, there are many harms due to models *not being capable enough*, for example jailbreaking [44] occurs because the models do not “understand” they are jailbroken. Collaborating with AI can help find research solutions to fix these problems – their own problems! – i.e., find and implement new capabilities that lead to safer models, new safety procedures, and co-developing values, constraints, and constitutions. This optimistic viewpoint contends that AIs increased capabilities can thus be leveraged to decrease harms – if done correctly.

As AI become more capable there is also an optimistic opportunity to help with many other societal issues beyond their own impact. Rather than the self-improved superintelligence dystopian paradigm where an AI overlord dictates best practices to humans, the co-improvement paradigm suggests collaborative help to synthesize consensus and find actionable steps to solve problems. Multi-human and AI collaboration could help aggregate viewpoints, structure debate and help humans come to positive conclusions and outcomes.

Co-improvement and openness If humanity wants to improve its scientific knowledge, the clearest way to do that is to use the scientific method. This means conducting reproducible science with openly disseminated results, so others can verify or build on them, and collective knowledge can advance. Co-improvement can help this knowledge advance more quickly in both the fields of AI, and potentially any other field of science as well. We note there is a current shift away from open AI research from a number of industrial labs. As stated in [45] we agree that “*concerns about misuse should not be used as an excuse to reduce openness to a greater extent than is required, for instance, when the real motivation is about corporate competitiveness.*” Nevertheless in AI, as in other fields of science, we believe *managed openness* should be considered in order to combat societal harms when required [45], which should be an ongoing discussion as capabilities increase.

Relation to other existing positions

Related Positions Our position is related to treatise on human-centered AI [46, 47, 48, 49], but in our case more specifically for the goal of achieving (co-)superintelligence via collaborative research. Similarly, the work of [50] advocates for cooperative AI, and finding common ground with machines, where research would be a special case. Nevertheless, we agree that work on all forms of human-AI collaboration is also an important goal for future human society with the advent of superhuman AIs. Other works emphasize that misalignment is a challenging goal, and advocate for the importance of AI’s objectives to be human-oriented [51]. Importantly, our position is that this challenging problem can and should be tackled by collaboration. For example,

collaboration with AI may more readily help find flaws in their own designs; however, we believe this research should be in progress now – not added on later when systems are fully deployed and it is too late.

Contrasting Positions Various works have described autonomous self-improvement and possible ways it can be achieved, for example [9, 52, 53, 54, 55]. Correspondingly, a large number of works have also developed practical instantiations with various technical contributions, see Table 2 for examples.

The authors of [54] advocate for “an era of experience” where self-improvement is obtained by autonomous learning from the AI’s own experience. The implication is that there is little cooperation with humans, for example they write that AIs will “*autonomously design and conduct experiments in fields like materials science, medicine, or hardware design*”. They also admit that this “*provides fewer opportunities for humans to intervene and mediate the agent’s actions, and therefore requires a high bar of trust and responsibility*”. Others see little role for humans when the self-improvement goal is achieved, e.g. from [56]: “... *AI will colonise the galaxy. Humans are not going to play a big role there, but that’s ok. We should be proud of being part of a grand process that transcends humankind.*”

In contrast, we envision a world where humans are always a necessary, but maximally augmented, part of not only economic, scientific, but also all types of decision-making processes. We believe the AI community should fully embrace and implement this vision in our own pursuit of this long term goal.

Conclusion

We have argued that the existing goal of autonomous self-improving AI is misguided for two reasons: it is not the fastest way to achieve superintelligence, nor the safest. We advocate instead for co-improvement, whereby we (human researchers) focus on building AI that is collaborative and in particular collaborates with us to conduct research – to help us achieve ever more collaborative, capable and safer AI, with its help. Achieving this goal in turn could unlock the creation of extremely capable AIs in the future that can collaborate with us to solve the important goals and societal problems for humanity at large.

Acknowledgements

This position paper, like the co-improvement paradigm we advocate for, was made with the help of AI. We also thank the humans at the AIRA and RAM teams in FAIR for useful discussions.

References

- [1] Alan M Turing. Computing machinery and intelligence (1950). *Mind*, 59(236):33–60, 2021.
- [2] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
- [3] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [4] Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. Self-challenging language model agents. *arXiv preprint arXiv:2506.01716*, 2025.
- [5] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.
- [6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [7] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

- [8] Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. [arXiv preprint arXiv:2408.02666](#), 2024.
- [9] Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In [Artificial general intelligence](#), pages 199–226. Springer, 2007.
- [10] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. [ACM Computing Surveys \(CSUR\)](#), 54(4):1–34, 2021.
- [11] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. [arXiv preprint arXiv:2408.06292](#), 2024.
- [12] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai’s ability to replicate ai research. [arXiv preprint arXiv:2504.01848](#), 2025.
- [13] Alexis Audran-Reiss, Jordi Armengol EstapÃ, Karen Hambardzumyan, Amar Budhiraja, Martin Josifoski, Edan Toledo, Rishi Hazra, Despoina Magka, Michael Shvartsman, Parth Pathak, et al. What does it take to be a good ai research agent? studying the role of ideation diversity. [arXiv preprint arXiv:2511.15593](#), 2025.
- [14] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. Mlgym: A new framework and benchmark for advancing ai research agents. [arXiv preprint arXiv:2502.14499](#), 2025.
- [15] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. [arXiv preprint arXiv:2306.12001](#), 2023.
- [16] Jiessie Tie, Bingsheng Yao, Tianshi Li, Syed Ishtiaque Ahmed, Dakuo Wang, and Shurui Zhou. Llms are imperfect, then what? an empirical study on llm failures in software engineering. [arXiv preprint arXiv:2411.09916](#), 2024.
- [17] Richard S Sutton, Andrew G Barto, et al. [Reinforcement learning: An introduction](#), volume 1. MIT press Cambridge, 1998.
- [18] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, et al. Playing atari with deep reinforcement learning. [arXiv:1312.5602](#), 2013.
- [19] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. [nature](#), 550(7676):354–359, 2017.
- [20] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. [Science](#), 362(6419):1140–1144, 2018.
- [21] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. [nature](#), 575(7782):350–354, 2019.
- [22] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. [arXiv preprint arXiv:1912.06680](#), 2019.
- [23] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. [arXiv preprint arXiv:2401.08967](#), 2024.
- [24] Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. [Advances in Neural Information Processing Systems](#), 37:116617–116637, 2024.
- [25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. [arXiv preprint arXiv:2501.12948](#), 2025.
- [26] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In [Advances in Neural Information Processing Systems](#), 2022.

- [27] Ping Yu, Jack Lanchantin, Tianlu Wang, Weizhe Yuan, Olga Golovneva, Ilia Kulikov, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. Cot-self-instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks. arXiv preprint arXiv:2507.23751, 2025.
- [28] Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. arXiv preprint arXiv:2510.24684, 2025.
- [29] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.
- [30] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. Journal of Machine Learning Research, 20(55):1–21, 2019.
- [31] Alexander Novikov, Ngăn Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131, 2025.
- [32] Eliezer Yudkowsky et al. Artificial intelligence as a positive and negative factor in global risk. Global catastrophic risks, 1(303):184, 2008.
- [33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- [36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.
- [37] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [38] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.
- [39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
- [40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.
- [41] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.
- [42] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874, 2021.
- [43] Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. arXiv preprint arXiv:2406.11931, 2024.
- [44] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? Advances in Neural Information Processing Systems, 36:80079–80110, 2023.
- [45] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228, 2018.

- [46] Ben Shneiderman. Human-centered AI. Oxford University Press, 2022.
- [47] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. arXiv preprint arXiv:2005.00582, 2020.
- [48] Eric J Horvitz. Reflections on challenges and promises of mixed-initiative interaction. AI Magazine, 28(2):3–3, 2007.
- [49] Mark Zuckerberg. Personal superintelligence for everyone. Blog post / open letter on Meta website, Jul 2025.
- [50] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. Nature, 593(7857):33–36, 2021.
- [51] Stuart Russell. Human-compatible artificial intelligence. Human-like machine intelligence, 1:3–22, 2022.
- [52] Jeff Clune. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. arXiv preprint arXiv:1905.10985, 2019.
- [53] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. Artificial intelligence, 299:103535, 2021.
- [54] David Silver and Richard S. Sutton. Welcome to the era of experience. In Designing an Intelligence. MIT Press, 2025. Preprint.
- [55] Nick Bostrom. Superintelligence. Dunod, 2024.
- [56] Jürgen Schmidhuber. Ai isn’t coming for our jobs, it’s coming for our planet and will one day ‘colonise the galaxy’. WIRED.