Welcome to the Era of Experience

David Silver, Richard S. Sutton*

Abstract

We stand on the threshold of a new era in artificial intelligence that promises to achieve an unprecedented level of ability. A new generation of agents will acquire superhuman capabilities by learning predominantly from experience. This note explores the key characteristics that will define this upcoming era.

The Era of Human Data

Artificial intelligence (AI) has made remarkable strides over recent years by training on massive amounts of human-generated data and fine-tuning with expert human examples and preferences. This approach is exemplified by large language models (LLMs) that have achieved a sweeping level of generality. A single LLM can now perform tasks spanning from writing poetry and solving physics problems to diagnosing medical issues and summarising legal documents.

However, while imitating humans is enough to reproduce many human capabilities to a competent level, this approach in isolation has not and likely cannot achieve superhuman intelligence across many important topics and tasks. In key domains such as mathematics, coding, and science, the knowledge extracted from human data is rapidly approaching a limit. The majority of high-quality data sources - those that can actually improve a strong agent's performance - have either already been, or soon will be consumed. The pace of progress driven solely by supervised learning from human data is demonstrably slowing, signalling the need for a new approach. Furthermore, valuable new insights, such as new theorems, technologies or scientific breakthroughs, lie beyond the current boundaries of human understanding and cannot be captured by existing human data.

The Era of Experience

To progress significantly further, a new source of data is required. This data must be generated in a way that continually improves as the agent becomes stronger; any static procedure for synthetically generating data will quickly become outstripped. This can be achieved by allowing agents to learn continually from their own *experience*, i.e., data that is generated by the agent interacting with its environment. AI is at the cusp of a new period in which experience will become the dominant medium of improvement and ultimately dwarf the scale of human data used in today's systems.

This transition may have already started, even for the large language models that epitomise human-centric AI. One example is in the capability of mathematics. AlphaProof [20] recently became the first program to achieve a medal in the International Mathematical Olympiad, eclipsing the performance of human-centric approaches [27, 19]. Initially exposed to around a hundred thousand formal proofs, created over many years

*This is a preprint of a chapter that will appear in the book *Designing an Intelligence*, published by MIT Press.

by human mathematicians, AlphaProof's reinforcement learning (RL) algorithm[1] subsequently generated a hundred million more through continual interaction with a formal proving system. This focus on interactive experience allowed AlphaProof to explore mathematical possibilities beyond the confines of pre-existing formal proofs, so as to discover solutions to novel and challenging problems. Informal mathematics has also achieved success by replacing expert generated data with self-generated data; for example, recent work from DeepSeek "underscores the power and beauty of reinforcement learning: rather than explicitly teaching the model on how to solve a problem, we simply provide it with the right incentives, and it autonomously develops advanced problem-solving strategies." [10]

Our contention is that incredible new capabilities will arise once the full potential of experiential learning is harnessed. This era of experience will likely be characterised by agents and environments that, in addition to learning from vast quantities of experiential data, will break through the limitations of human-centric AI systems in several further dimensions:

- Agents will inhabit streams of experience, rather than short snippets of interaction.
- Their actions and observations will be richly grounded in the environment, rather than interacting via human dialogue alone.
- Their rewards will be grounded in their experience of the environment, rather than coming from human prejudgement.
- They will plan and/or reason about experience, rather than reasoning solely in human terms

We believe that today's technology, with appropriately chosen algorithms, already provides a sufficiently powerful foundation to achieve these breakthroughs. Furthermore, the pursuit of this agenda by the AI community will spur new innovations in these directions that rapidly progress AI towards truly superhuman agents.

Streams

An experiential agent can continue to learn throughout a lifetime. In the era of human data, language-based AI has largely focused on short interaction episodes: e.g., a user asks a question and (perhaps after a few thinking steps or tool-use actions) the agent responds. Typically, little or no information carries over from one episode to the next, precluding any adaptation over time. Furthermore, the agent aims exclusively for outcomes within the current episode, such as directly answering a user's question. In contrast, humans (and other animals) exist in an ongoing stream of actions and observations that continues for many years. Information is carried across the entire stream, and their behaviour adapts from past experiences to self-correct and improve. Furthermore, goals may be specified in terms of actions and observations that stretch far into the future of the stream. For example, humans may select actions to achieve long-term goals like improving their health, learning a language, or achieving a scientific breakthrough.

Powerful agents should have their own stream of experience that progresses, like humans, over a long time-scale. This will allow agents to take actions to achieve future goals, and to continuously adapt over time to new patterns of behaviour. For example, a health and wellness agent connected to a user's wearables could monitor sleep patterns, activity levels, and dietary habits over many months. It could then provide personalized recommendations, encouragement, and adjust its guidance based on long-term trends and the user's specific health goals. Similarly, a personalized education agent could track a user's progress in learning

[1] An RL algorithm is one that learns to achieve a goal by trial and error, i.e., adapting its behaviour from its experience of interacting with its environment. Adaptation may happen by any means, for example updating the weights of a neural network, or adapting in-context based on feedback from the environment.

a new language, identify knowledge gaps, adapt to their learning style, and adjust its teaching methods over months or even years. Furthermore, a science agent could pursue ambitious goals, such as discovering a new material or reducing carbon dioxide. Such an agent could analyse real-world observations over an extended period, developing and running simulations, and suggesting real-world experiments or interventions.

In each case, the agent takes a sequence of steps so as to maximise long-term success with respect to the specified goal. An individual step may not provide any immediate benefit, or may even be detrimental in the short term, but may nevertheless contribute in aggregate to longer term success. This contrasts strongly with current AI systems that provide immediate responses to requests, without any ability to measure or optimise the future consequences of their actions on the environment.

Actions and Observations

Agents in the era of experience will act autonomously in the real world. LLMs in the era of human data focused primarily on human-privileged actions and observations that output text to a user, and input text from the user back into the agent. This differs markedly from natural intelligence, in which an animal interacts with its environment through motor control and sensors. While animals, and most notably humans, may communicate with other animals, this occurs through the same interface as other sensorimotor control rather than a privileged channel.

It has long been recognised that LLMs may also invoke actions in the digital world, for example by calling APIs (see for example, [43]). Initially, these capabilities came largely from human examples of tool-use, rather than from the experience of the agent. However, coding and tool-use capabilities have built increasingly upon execution feedback [17, 7, 12], where the agent actually runs code and observes what happens. Recently, a new wave of prototype agents have started to interact with computers in an even more general manner, by using the same interface that humans use to operate a computer [3, 15, 24]. These changes herald a transition from exclusively human-privileged communication, to much more autonomous interactions where the agent is able to act independently in the world. Such agents will be able to actively explore the world, adapt to changing environments, and discover strategies that might never occur to a human.

These richer interactions will provide a means to autonomously understand and control the digital world. The agent may use 'human-friendly' actions and observations such as user interfaces, that naturally facilitate communication and collaboration with the user. The agent may also take 'machine-friendly' actions that execute code and call APIs, allowing the agent to act autonomously in service of its goals. In the era of experience, agents will also interact with the real world via digital interfaces. For example, a scientific agent could monitor environmental sensors, remotely operate a telescope, or control a robotic arm in a laboratory to autonomously conduct experiments.

Rewards

What if experiential agents could learn from external events and signals, and not just human preferences?

Human-centric LLMs typically optimise for rewards based on human prejudgement: an expert observes the agent's action and decides whether it is a good action, or picks the best agent action among multiple alternatives. For example, an expert may judge a health agent's advice, an educational assistant's teaching, or a scientist agent's suggested experiment. The fact that these rewards or preferences are determined by humans in absence of their consequences, rather than measuring the effect of those actions on the environment, means that they are not directly grounded in the reality of the world. Relying on human prejudgement in this manner usually leads to an impenetrable ceiling on the agent's performance: the agent cannot discover better strategies that are underappreciated by the human rater. To discover new ideas that go far beyond existing human knowledge, it is instead necessary to use grounded rewards: signals that arise from the environment itself. For example, a health assistant could ground the user's health goals into a reward based on a combination of

signals such as their resting heart rate, sleep duration, and activity levels, while an educational assistant could use exam results to provide a grounded reward for language learning. Similarly, a science agent with a goal to reduce global warming might use a reward based on empirical observations of carbon dioxide levels, while a goal to discover a stronger material might be grounded in a combination of measurements from a materials simulator, such as tensile strength or Young's modulus.

Grounded rewards may arise from humans that are part of the agent's environment.[2] For example, a human user could report whether they found a cake tasty, how fatigued they are after exercising, or the level of pain from a headache, enabling an assistant agent to provide better recipes, refine its fitness suggestions, or improve its recommended medication. Such rewards measure the consequence of the agent's actions within their environment, and should ultimately lead to better assistance than a human expert that prejudges a proposed cake recipe, exercise program, or treatment program.

Where do rewards come from, if not from human data? Once agents become connected to the world through rich action and observation spaces (see above), there will be no shortage of grounded signals to provide a basis for reward. In fact, the world abounds with quantities such as cost, error rates, hunger, productivity, health metrics, climate metrics, profit, sales, exam results, success, visits, yields, stocks, likes, income, pleasure/pain, economic indicators, accuracy, power, distance, speed, efficiency, or energy consumption. In addition there are innumerable additional signals arising from the occurrence of specific events, or from features derived from raw sequences of observations and actions.

One could in principle create a variety of distinct agents, each optimising for one grounded signal as its reward. There is an argument that even a single such reward signal, optimised with great effectiveness, may be sufficient to induce broadly capable intelligence [34].[3] This is because the achievement of a simple goal in a complex environment may often require a wide variety of skills to be mastered.

However, the pursuit of a single reward signal does not on the surface appear to meet the requirements of a general-purpose AI that can be steered reliably towards arbitrary user-desired behaviours. Is the autonomous optimisation of grounded, non-human reward signals therefore in opposition to the requirements of modern AI systems? We argue that this is not necessarily the case, by sketching one approach that may meet these desiderata; other approaches may also be possible.

The idea is to flexibly adapt the reward, based on grounded signals, in a user-guided manner. For example, the reward function could be defined by a neural network that takes the agent's interactions with both the user and the environment as input, and outputs a scalar reward. This allows the reward to select or combine together signals from the environment in a manner that depends upon the user's goal. For example, a user might specify a broad goal such as 'improve my fitness' and the reward function might return a function of the user's heart rate, sleep duration, and steps taken. Or the user might specify a goal of 'help me learn Spanish' and the reward function could return the user's Spanish exam results.

Furthermore, users could provide feedback during the learning process, such as their satisfaction level, which could be used to fine-tune the reward function. The reward function can then adapt over time, to improve the way in which it selects or combines signals, and to identify and correct any misalignment. This can also be understood as a bi-level optimisation process that optimises user feedback as the top-level goal, and optimises grounded signals from the environment at the low level.[4] In this way, a small amount of human data may facilitate a large amount of autonomous learning.

[2] Experience and human data are not exact opposites. For example, a dog learns entirely from experience, but human interaction is part of its experience.

[3] The reward-is-enough hypothesis suggests that intelligence, and its associated abilities, can emerge naturally from the maximisation of reward. This may include environments containing human interaction and rewards based on human feedback.

[4] In this case, one may also view grounded human feedback as a singular reward function forming the agent's overall objective, which is maximised by constructing and optimising an intrinsic reward function [8] based on rich, grounded feedback.

Planning and Reasoning

Will the era of experience change the way that agents plan and reason? Recently, there has been significant progress using LLMs that can reason, or "think" with language [23, 14, 10], by following a chain of thought before outputting a response [16]. Conceptually, LLMs can act as a universal computer [30]: an LLM can append tokens into its own context, allowing it to execute arbitrary algorithms before outputting a final result.

In the era of human data, these reasoning methods have been explicitly designed to imitate human thought processes. For example, LLMs have been prompted to emit human-like chains of thought [16], imitate traces of human thinking [42], or to reinforce steps of thinking that match human examples [18]. The reasoning process may be fine-tuned further to produce thinking traces that match the correct answer, as determined by human experts [44].

However, it is highly unlikely that human language provides the optimal instance of a universal computer. More efficient mechanisms of thought surely exist, using non-human languages that may for example utilise symbolic, distributed, continuous, or differentiable computations. A self-learning system can in principle discover or improve such approaches by learning how to think from experience. For example, AlphaProof learned to formally prove complex theorems in a manner quite different to human mathematicians [20].

Furthermore, the principle of a universal computer only addresses the internal computation of the agent; it does not connect it to the realities of the external world. An agent trained to imitate human thoughts or even to match human expert answers may inherit fallacious methods of thought deeply embedded within that data, such as flawed assumptions or inherent biases. For example, if an agent had been trained to reason using human thoughts and expert answers from 5,000 years ago it may have reasoned about a physical problem in terms of animism; 1,000 years ago it may have reasoned in theistic terms; 300 years ago it may have reasoned in terms of Newtonian mechanics; and 50 years ago in terms of quantum mechanics. Progressing beyond each method of thought required interaction with the real world: making hypotheses, running experiments, observing results, and updating principles accordingly. Similarly, an agent must be grounded in real-world data in order to overturn fallacious methods of thought. This grounding provides a feedback loop, allowing the agent to test its inherited assumptions against reality and discover new principles that are not limited by current, dominant modes of human thought. Without this grounding, an agent, no matter how sophisticated, will become an echo chamber of existing human knowledge. To move beyond this, agents must actively engage with the world, collect observational data, and use that data to iteratively refine their understanding, mirroring in many ways the process that has driven human scientific progress.

One possible way to directly ground thinking in the external world is to build a world model [37] that predicts the consequences of the agent's actions upon the world, including predicting reward. For example, a health assistant might consider making a recommendation for a local gym or a health podcast. The agent's world model might predict how a user's heart rate or sleep patterns might subsequently change following this action, as well as predicting future dialogue with the user. This allows the agent to plan [36, 29] directly in terms of its own actions and their causal effect upon the world. As the agent continues to interact with the world throughout its stream of experience, its dynamics model is continually updated to correct any errors in its predictions. Given a world model, an agent may apply scalable planning methods that improve the predicted performance of the agent.

Planning and reasoning methods are not mutually exclusive: an agent may apply internal LLM computations to select each action during planning, or to simulate and evaluate the consequences of those actions.

Why Now?

Learning from experience is not new. Reinforcement learning systems have previously mastered a large number of complex tasks that were represented in a simulator with a clear reward signal (c.f., approximately, the "era of simulation" in Figure 1). For example, RL methods equalled or exceeded human performance

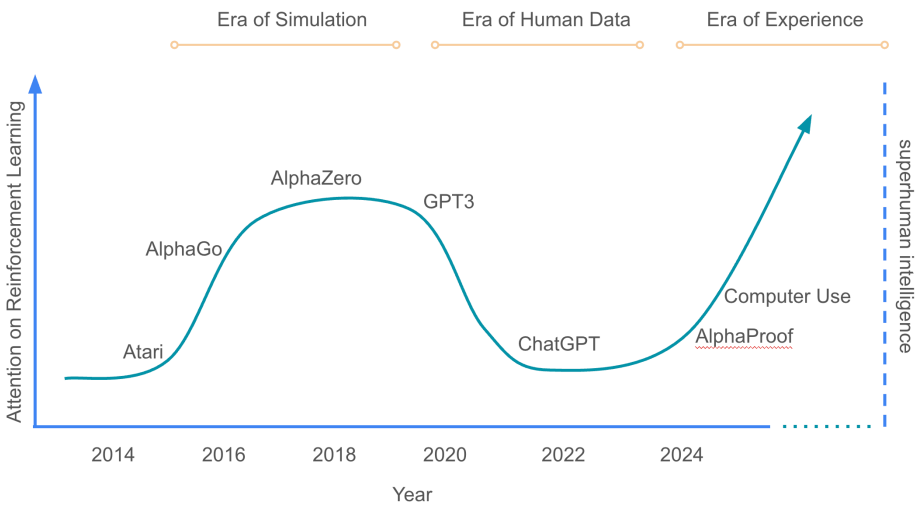

Figure 1: A sketch chronology of dominant AI paradigms. The y-axis suggests the proportion of the field's total effort and computation that is focused on RL.

through self-play in board games such as backgammon [39], Go [31], chess [32], poker [22, 6] and Stratego [26]; video games such as Atari [21], StarCraft II [40], Dota 2 [4] and Gran Turismo [41]; dextrous manipulation tasks such as Rubik's cube [1]; and resource management tasks such as data center cooling [13]. Furthermore, powerful RL agents such as AlphaZero [33] exhibited impressive and potentially unlimited scalability with the size of the neural network, the quantity of interactive experience, and the duration of thinking time. However, agents based on this paradigm did not leap the gap between simulation (closed problems with singular, precisely defined rewards) to reality (open-ended problems with a plurality of seemingly ill-defined rewards).

The era of human data offered an appealing solution. Massive corpuses of human data contain examples of natural language for a huge diversity of tasks. Agents trained on this data achieved a wide range of competencies compared to the more narrow successes of the era of simulation. Consequently, the methodology of experiential RL was largely discarded in favour of more general-purpose agents, resulting in a widespread transition to human-centric AI.

However, something was lost in this transition: an agent's ability to self-discover its own knowledge. For example, AlphaZero discovered fundamentally new strategies for chess and Go, changing the way that humans play these games [28, 45]. The era of experience will reconcile this ability with the level of task-generality achieved in the era of human data. This will become possible, as outlined above, when agents are able to autonomously act and observe in streams of real-world experience [11], and where the rewards may be flexibly connected to any of an abundance of grounded, real-world signals. The advent of autonomous agents that interact with complex, real-world action spaces [3, 15, 24], alongside powerful RL methods that can solve open-ended problems in rich reasoning spaces [20, 10] suggests that the transition to the era of experience is imminent.

Reinforcement Learning Methods

Reinforcement learning (RL) has a rich history that is deeply rooted in autonomous learning, where agents learn for themselves through direct interaction with their environment. Early RL research yielded a suite of powerful concepts and algorithms. For example, temporal difference learning [35] enabled agents to estimate future rewards, leading to breakthroughs such as superhuman performance in backgammon [39]. Exploration techniques, driven by optimism or curiosity, were developed to help agents discover creative new behaviors and avoid getting stuck in suboptimal routines [2]. Methods like the Dyna algorithm enabled agents to build and learn from models of their world, allowing them to plan and reason about future actions [36, 29]. Concepts like options and inter/intra-option learning facilitated temporal abstraction, enabling agents to reason over longer timescales and break down complex tasks into manageable sub-goals [38].

The rise of human-centric LLMs, however, shifted the focus away from autonomous learning and towards leveraging human knowledge. Techniques like RLHF (Reinforcement Learning from Human Feedback) [9, 25] and methods for aligning language models with human reasoning [44] proved incredibly effective, driving rapid progress in AI capabilities. These approaches, while powerful, often bypassed core RL concepts: RLHF side-stepped the need for value functions by invoking human experts in place of machine-estimated values, strong priors from human data reduced the reliance on exploration, and reasoning in human-centric terms lessened the need for world models and temporal abstraction.

However, it could be argued that the shift in paradigm has thrown out the baby with the bathwater. While human-centric RL has enabled an unprecedented breadth of behaviours, it has also imposed a new ceiling on the agent's performance: agents cannot go beyond existing human knowledge. Furthermore, the era of human data has focused predominantly on RL methods that are designed for short episodes of ungrounded, human interaction, and are not suitable for long streams of grounded, autonomous interaction.

The era of experience presents an opportunity to revisit and improve classic RL concepts. This era will bring new ways to think about reward functions that are flexibly grounded in observational data. It will revisit value functions and methods to estimate them from long streams with as yet incomplete sequences. It will bring principled yet practical methods for real-world exploration that discover new behaviours that are radically different from human priors. Novel approaches to world models will be developed that capture the complexities of grounded interactions. New methods for temporal abstraction will allow agents to reason, in terms of experience, over ever-longer time horizons. By building upon the foundations of RL and adapting its core principles to the challenges of this new era, we can unlock the full potential of autonomous learning and pave the way to truly superhuman intelligence.

Consequences

The advent of the era of experience, where AI agents learn from their interactions with the world, promises a future profoundly different from anything we have seen before. This new paradigm, while offering immense potential, also presents important risks and challenges that demand careful consideration, including but not limited to the following points.

On the positive side, experiential learning will unlock unprecedented capabilities. In everyday life, personalized assistants will leverage continuous streams of experience to adapt to individuals' health, educational, or professional needs towards long-term goals over the course of months or years. Perhaps most transformative will be the acceleration of scientific discovery. AI agents will autonomously design and conduct experiments in fields like materials science, medicine, or hardware design. By continuously learning from the results of their own experiments, these agents could rapidly explore new frontiers of knowledge, leading to the development of novel materials, drugs, and technologies at an unprecedented pace.

However, this new era also presents significant and novel challenges. While the automation of human

capabilities promises to boost productivity, these improvements could also lead to job displacement. Agents may even be able to exhibit capabilities previously considered the exclusive realm of humanity, such as long-term problem-solving, innovation, and a deep understanding of real world consequences.

Furthermore, whilst general concerns exist around the potential misuse of any AI, heightened risks may arise from agents that can autonomously interact with the world over extended periods of time to achieve long-term goals. By default, this provides fewer opportunities for humans to intervene and mediate the agent's actions, and therefore requires a high bar of trust and responsibility. Moving away from human data and human modes of thinking may also make future AI systems harder to interpret.

However, whilst acknowledging that experiential learning will increase certain safety risks, and that further research is surely required to ensure a safe transition into the era of experience, we should also recognise that it may also provide some important safety benefits.

Firstly, an experiential agent is aware of the environment it is situated within, and its behaviour can adapt over time to changes in that environment. Any pre-programmed system, including a fixed AI system, can be unaware of its environmental context, and become maladapted to the changing world into which it is deployed. For example, a critical piece of hardware may malfunction, a pandemic might cause rapid societal change, or a new scientific discovery may trigger a cascade of rapid technological developments. By contrast, an experiential agent could observe and learn to circumvent malfunctioning hardware, adjust to rapid societal change, or embrace and build upon new science and technology. Perhaps even more importantly, the agent could recognise when its behaviour is triggering human concern, dissatisfaction, or distress, and adaptively modify its behaviour to avoid these negative consequences.

Secondly, the agent's reward function may itself be adapted through experience, for example using the bi-level optimisation described earlier (see Rewards). Importantly, this means that misaligned reward functions can often be incrementally corrected over time by trial and error. For example, rather than blindly optimising a signal, such as the maximisation of paperclips [5], the reward function could be modified, based upon indications of human concern, before paperclip production consumes all of the Earth's resources. This is analogous to the way that humans set goals for each other, and then adapt those goals if they observe people gaming the system, neglecting long-term well-being, or causing undesired negative consequences; although also like human goal-setting, there is no guarantee of perfect alignment.

Finally, advancements relying on physical experience are inherently constrained by the time it takes to execute actions in the real world and observe their consequences. For example, the development of a new drug, even with AI-assisted design, still requires real-world trials that cannot be completed overnight. This may provide a natural brake on the pace of potential AI self-improvement.

Conclusion

The era of experience marks a pivotal moment in the evolution of AI. Building on today's strong foundations, but moving beyond the limitations of human-derived data, agents will increasingly learn from their own interactions with the world. Agents will autonomously interact with environments through rich observations and actions. They will continue to adapt over the course of lifelong streams of experience. Their goals will be directable towards any combination of grounded signals. Furthermore, agents will utilise powerful non-human reasoning, and construct plans that are grounded in the consequences of the agent's actions upon its environment. Ultimately, experiential data will eclipse the scale and quality of human generated data. This paradigm shift, accompanied by algorithmic advancements in RL, will unlock in many domains new capabilities that surpass those possessed by any human.

Acknowledgements

The authors would like to acknowledge helpful comments and discussion from Thomas Degris, Rohin Shah, Tom Schaul and Hado van Hasselt.

References

[1] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving Rubik's cube with a robot hand, 2019.

[2] S. Amin, M. Gomrokchi, H. Satija, H. van Hoof, and D. Precup. A survey of exploration methods in reinforcement learning, 2021.

[3] Anthropic. Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. `https://www.anthropic.com/news/3-5-models-and-computer-use`, 2024.

[4] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang. Dota 2 with large scale deep reinforcement learning, 2019.

[5] N. Bostrom. Ethical issues in advanced artificial intelligence. `https://nickbostrom.com/ethics/ai`, 2003.

[6] N. Brown and T. Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[7] X. Chen, M. Lin, N. Schärli, and D. Zhou. Teaching large language models to self-debug, 2023.

[8] N. Chentanez, A. Barto, and S. Singh. Intrinsically motivated reinforcement learning. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

[9] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] DeepSeek AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[11] M. Elsayed, G. Vasan, and A. R. Mahmood. Streaming deep reinforcement learning finally works, 2024.

[12] J. Gehring, K. Zheng, J. Copet, V. Mella, Q. Carbonneaux, T. Cohen, and G. Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning, 2025.

[13] Google DeepMind. Deepmind AI reduces google data centre cooling bill by 40%. `https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/`, 2016.

[14] Google DeepMind. Gemini: Flash thinking. `https://deepmind.google/technologies/gemini/flash-thinking/`, 2024.

[15] Google DeepMind. Project Mariner. `https://deepmind.google/technologies/project-mariner`, 2024.

[16] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022.

[17] H. Le, Y. Wang, A. D. Gotmare, S. Savarese, and S. C. H. Hoi. CodeRL: Mastering code generation through pretrained models and deep reinforcement learning, 2022.

[18] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step, 2023.

[19] H. Mahdavi, A. Hashemi, M. Daliri, P. Mohammadipour, A. Farhadi, S. Malek, Y. Yazdanifard, A. Khasahmadi, and V. Honavar. Brains vs. bytes: Evaluating llm proficiency in olympiad mathematics, 2025.

[20] H. Masoom, A. Huang, M. Z. Horváth, T. Zahavy, V. Veeriah, E. Wieser, J. Yung, L. Yu, Y. Schroecker, J. Schrittwieser, O. Bertolli, B. Ibarz, E. Lockhart, E. Hughes, M. Rowland, G. Margand, A. Davies, D. Zheng, I. Beloshapka, I. von Glehn, Y. Li, F. Pedregosa, A. Velingker, G. Žužić, O. Nash, B. Mehta, P. Lezeau, S. Mercuri, L. Wu, C. Soenne, T. Murrills, L. Massacci, A. Yang, A. Mandhane, T. Eccles, E. Aygün, Z. Gong, R. Evans, S. Mokrá, A. Barekatain, W. Shang, H. Openshaw, F. Gimeno, D. Silver, and P. Kohli. AI achieves silver-medal standard solving International Mathematical Olympiad problems. `https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/`, 2024.

[21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[22] M. Moravčík, M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[23] OpenAI. Openai o1 mini: Advancing cost-efficient reasoning. `https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/`, 2024.

[24] OpenAI. Introducing Operator. `https://openai.com/index/introducing-operator`, 2025.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.

[26] J. Perolat, B. D. Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, S. McAleer, R. Elie, S. H. Cen, Z. Wang, A. Gruslys, A. Malysheva, M. Khan, S. Ozair, F. Timbers, T. Pohlen, T. Eccles, M. Rowland, M. Lanctot, J.-B. Lespiau, B. Piot, S. Omidshafiei, E. Lockhart, L. Sifre, N. Beauguerlange, R. Munos, D. Silver, S. Singh, D. Hassabis, and K. Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

[27] I. Petrov, J. Dekoninck, L. Baltadzhiev, M. Drencheva, K. Minchev, M. Balunović, N. Jovanović, and M. Vechev. Proof or bluff? evaluating llms on 2025 usa math olympiad, 2025.

[28] M. Sadler and N. Regan. *Game Changer*. New in Chess, 2019.

[29] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. P. Lillicrap, and D. Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604 – 609, 2019.

[30] D. Schurmanns. Memory augmented large language models are computationally universal. *arXiv preprint arXiv:2501.12948*, 2023.

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

[32] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

[33] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Grapel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[34] D. Silver, S. Singh, D. Precup, and R. S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[35] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[36] R. S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224. Morgan Kaufmann, 1990.

[37] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[38] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

[39] G. Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219, 1994.

[40] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. P. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350 – 354, 2019.

[41] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, P. Khandelwal, V. Kompella, H. Lin, P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, and H. Kitano. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[42] M. S. Yang, D. Schuurmans, P. Abbeel, and O. Nachum. Chain of thought imitation with procedure cloning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36366–36381. Curran Associates, Inc., 2022.

[43] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in large language models. In *11th International Conference on Learning Representations*, 2023.

[44] E. Zelikman, J. M. Mu, N. D. Goodman, and G. Poesia. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:24170–24184, 2022.

[45] Y. Zhou. *Rethinking Opening Strategy: AlphaGo's Impact on Pro Play*. CreateSpace Independent, 2018.