

AI Research Paper

Zoe Mail

4.14.24

Digital Revolution

To go about this paper, I debated between a creative route compared to a more traditional one. Ultimately I went the route of writing a traditional paper, as I felt I would learn more from this in a way that would apply to future endeavors of mine. I had used ChatGPT a few times in the past, but not regularly enough or in an experimental way to feel like I had a strong grasp of the platform or other AI platforms. At the start, I had a general idea of the prompts I wanted to use but throughout the process had to adapt to the answers I was given. The most important aspect to me of writing the paper was getting the tone right. In past experiences using AI, tone was the main issue with the responses I was receiving. Throughout this paper, all yellow highlighted sections are direct prompts I submitted in the chatbots.

Prompt 1, ChatGPT: describe the development of artificial intelligence from Turing to large language model chatbots:

The development of artificial intelligence (AI) from Turing to large language model chatbots represents a journey marked by significant milestones and advancements in the field of computer science and machine learning. Here's a brief overview:

1. ****Alan Turing and Early Concepts (1940s-1950s)**:**

- Alan Turing laid the groundwork for AI with his seminal paper "Computing Machinery and Intelligence" in 1950, which proposed the Turing Test as a measure of a machine's intelligence.

- Early AI research focused on symbolic reasoning, logical inference, and problem-solving.

2. ****Symbolic AI and Expert Systems (1960s-1980s)**:**

- Symbolic AI, also known as "good old-fashioned AI" (GOFAI), dominated the field during this period. Researchers developed expert systems that encoded human knowledge in a set of rules.

- Examples include programs like MYCIN for medical diagnosis and DENDRAL for chemical analysis.

3. ****Connectionism and Neural Networks (1980s-1990s)**:**

- Connectionism emerged as an alternative to symbolic AI, emphasizing the use of artificial neural networks inspired by the structure and function of the human brain.

- Neural networks, particularly backpropagation, gained attention for their ability to learn from data and adapt to complex patterns.

4. ****AI Winter and Rebirth (1990s-2000s)**:**

- Following over-hyped expectations and underwhelming results, the AI field experienced an "AI winter" during the late 20th century.

- However, advancements in areas like machine learning, natural language processing, and robotics laid the foundation for the resurgence of AI in the early 21st century.

5. ****Deep Learning and Big Data (2010s)**:**

- The proliferation of big data and improvements in computational power fueled the rise of deep learning, a subfield of machine learning that utilizes neural networks with many layers.
- Deep learning achieved breakthroughs in image recognition, speech recognition, and natural language processing, leading to applications like self-driving cars, virtual assistants, and recommendation systems.

6. ****Large Language Model Chatbots (2010s-Present)****:

- Large language model chatbots represent one of the latest advancements in AI, leveraging deep learning techniques, particularly transformers like OpenAI's GPT series.
- These chatbots, such as GPT-3, are capable of generating human-like text based on prompts, engaging in conversations, answering questions, and even assisting in various tasks.

Throughout this development trajectory, AI has evolved from rule-based systems to data-driven approaches, with an increasing emphasis on learning from large datasets and adapting to complex environments. Large language model chatbots stand as a testament to the progress made in natural language understanding and generation, bringing AI closer to human-like interaction and communication.

Prompt 2, ChatGPT: If I'm writing a paper with this prompt, how do you suggest I structure the paper?

ChatGPT provided me with 9 sections for the paper and suggested content to include in each. I decided to aim for 280 words per section which would bring me to around 2500 words for

the paper. At this point, I was trying to decide which AI would provide me with the highest quality content. Further, I wanted to ensure the writing style would be suitable for this assignment. I found a scholarly article online about technology, from a reputable website: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9926968/>. I felt that the writing style was what I was looking for to write this paper.

Prompt 3: I submitted the following prompt to Gemini, ChatGPT, and Claude:

Here is an example of a scholarly article: (insert page-long section of the article)

Please write me 250 words emulating the writing style for this prompt: discuss the early foundations of AI, including Alan Turing's contributions and the Dartmouth Conference. Provide a brief overview of the key developments and milestones in AI leading up to the present day. Highlight the major paradigms and approaches in AI during different time periods.

The prompt part of this is directly from the first content section of the paper that ChatGPT suggested. I decided to skip the introduction section, as I felt like the body paragraphs were most important to get right and this response would act as a sample for the majority of the paper.

Prompt 4: Claude had already included in-text citations. Before deciding which site to continue with, the next prompt I submitted was to ask Gemini and ChatGPT “What are your sources?”

Prompt 5: Both Gemini and ChatGPT did not respond with sources, so I edited the prompt to clarify. I changed my prompt to “What are your sources for the response you just gave me?”

Gemini responded by merely saying it was trained on publicly available data and did not give sources but offered to help me learn more if I wanted. ChatGPT responded by explaining that it did not have direct access to external sources or databases. I decided that since Claude provided in-text citations, this would be the best model to write a research paper with. I was satisfied with the content it gave me for the section I asked. I chose a couple of the in-text citations to put into Google to verify that they were actual sources; Google confirmed that they were. I checked the word count of the section that Claude wrote for me and it was around 280 words, which I deemed sufficient.

Prompt 6: The next day I continued getting content for the paper. My Claude window was still open from before, but I wanted to ensure that the AI would still generate the same type of content answer with the tone used for the section already written. I prompted the AI “Using the same tone of voice as before, and on the same topic, write 280 words which include the following: Provide an overview of the paper’s scope and objectives. Introduce the topic of artificial intelligence and its significance in modern society. Present the main argument or thesis statement”. This was the suggested content to include in the paper introduction, as stated by ChatGPT in its outline.

Prompt 7: I noticed that the content provided included first-person writing in the section, which I think of as to be avoided for a research paper. The words “we” and “our” were used on numerous occasions. I prompted Claude “Please change to get rid of any first-person perspective”. However, when I read the section I noticed that these words were still in it.

Prompt 8: I read through the original and revised versions to try to find any changes made based on my prompt, but did not spot any. I told Claude “Highlight any changes from the

answer you just gave me to the one before”. Claude responded merely by once again sending the same content back to me and said at the start of the response “Here is the overview with changes from the previous version highlighted”. Nothing was highlighted, so I decided that this would be something I would edit manually once my paper draft was complete.

Prompt 9: I put the rest of the sections with suggested talking points from ChatGPT into Claude together with the prompt: Please write 280 words for each section listed and include all the content points. Write in the same scholarly tone as above. Avoid using the first person.

The sections looked a little scant to me, so I checked the word count and saw that they were only around 180 words each; many of them were even less. The exception was the conclusion which was around 220 words. The total paper at this point was close to 1700 words and I was aiming for 2500.

Prompt 10: At this point, I wanted to expand the paper, so I did some research to find a credible source that I wanted to feed Claude. I found a 30-page research paper from the University of Washington entitled ‘The History of Artificial Intelligence’. I typed: “Use this source to expand each section by 100 words: <https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf>”.

Unfortunately, the response I received was that Claude is not capable of reading external sources. Having used AI very few times before this assignment, I was not aware of this limitation.

Prompt 11: I wanted to input my own source to help Claude add to the paper, but thought that the 30-page research paper was too long to copy and paste into the chatbot. I did more online

research and found a shorter source with information I deemed relevant to expand the paper:
<https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>.

The prompt I used was: “This is more info on AI to utilize: (insert copied section from the source). Expand each of the following sections by 100 words: (insert all content sections of my paper)”

I found that Claude kept referring to my source as “the source” and was more heavy on the new source relative to the rest of the paper, which I was not the biggest fan of. I decided to redo this part and rephrase my prompt to include the author’s name so that he would be directly cited.

Prompt 12: This is more info on AI to utilize from author Rockwell Anyoha. Add 100 words to each written section using some knowledge from your database and some from the new source.

The chatbot was confused and asked me to supply content, so I then copied and pasted the content again and clarified. Rather than expanding what I had written, Claude rewrote the sections and still did not make them longer. Once again, I rewrote the prompt.

Prompt 13: Redo and keep the original section written, while adding 100 words to make it longer.

Finally, the original writing remained the same and words were added. However, only around 50-75 words were added per section, which was less than I was looking for once again. I

went through and added the new sentences to my paper draft. The new source I had suggested was excluded.

Prompt 14: Keep the original sections written and add 100 words to each. This is content to reference: (insert pasted material).

I received an answer that the content was too long and I would need to pay for a Premium subscription. I revised my prompt to fit the word count by referencing the content by author, since I had sent it several prompts before. I only wanted 50 more words per section, but figured that by asking Claude to add 100 words, it would give me less than I asked for as it had done in the past.

Prompt 15: Keep the original sections written and add 100 words to each using the above content I sent from author Rockwell Anyoha.

Claude added around 50 words to each section as before. It kept the original content from before the previous time I asked for more to be added. The new sentences that were added provided different content from before, yet I noticed they often started with the same few words. This is something I decided to go back and manually edit later. Since I had asked that Claude include content I had pasted from the Harvard article, there were several quotes from the author. Yet when comparing it to the previous version, I noticed that Claude had merely put quotes around a sentence it had included last time and turned it into a quote from the author. When I double-checked with the article, the quotes were nowhere to be around. So, I went back and deleted anything related to the author, as I could not confirm its validity. Each section had added 1-2 quotations to already existent sentences and turned them into quotes from the author whose

article I had used as a reference for new information. I decided to manually add information from the Harvard article to my paper instead, as Claude had shown me it could not be trusted to do this.

Prompt 16: Cite all sources used from all responses in the Chicago Manual Style.

To verify the sources were real, I googled many of them, especially those with strange-looking links. Surprisingly, all the ones I googled were valid sources.

From there, I went through to make edits of my own. The Grammarly plugin I use gave me several suggestions. Most of them were from Grammarly Premium, which I do not have, but I could see in which parts the suggestions were made. I thought it was interesting that Grammarly would want to improve the grammar of artificial intelligence. All the Grammarly Premium suggested changes are in orange, and the edits I made are in green.

From working on this paper, I felt like using the combination of chatbots helped greatly. Starting with the outline and then using the suggested content and specifying a word count is what enabled me to have a structured paper. Finding the right tone is something I have struggled with in the past while using AI. Copying and pasting an excerpt from the type of paper I wanted to write allowed me to have the tone of voice I was looking for.

AI Research Paper

Introduction

The field of artificial intelligence (AI) has seen remarkable progress in recent decades, permeating numerous aspects of modern society. Since the release of ChatGPT, numerous platforms have popped up, enabling it to spread widely throughout Western society. AI systems can now match or exceed human capabilities across a wide range of tasks, from game-playing and pattern recognition to language processing and decision-making. As AI becomes increasingly sophisticated and ubiquitous, it raises profound questions about the nature of intelligence itself and its implications for the future (Russell and Norvig, 2010). Artificially intelligent robots in science fiction movies are what familiarized society with this concept, though people were unaware of what this would turn into in the future and how it would infiltrate Western society (Anyoha, 2017).

This paper aims to provide a comprehensive examination of the current state of AI technology and its projected trajectory. The paper analyzes the key drivers behind AI's rapid development, including increases in computational power, availability of big data, and advances in machine learning algorithms and neural network architectures. The core objective is to assess AI's potential impacts across diverse domains, both benefiting humanity through automation and augmentation, but also surfacing potential risks that must be carefully navigated (Bostrom, 2014). With technology this powerful and capable, the risks increase exponentially. AI's ability to cause harm is just as great as its ability to benefit society.

The central thesis is that while AI undoubtedly represents a powerful transformative force, its positive realization hinges on the ability to align AI systems with human values and

ethical principles from the outset. People must be proactive in shaping AI's development in a controlled, responsible manner to ensure it remains robustly beneficial and avoids perilous unintended consequences (Amodei et al., 2016). Concrete proposals will be outlined for establishing an AI governance framework and research agenda prioritizing safety, robustness, and value alignment alongside continued capabilities gains. Having a governance framework is a necessary aspect of utilizing AI to keep it in check and ensure that moral boundaries are not crossed.

In an era of rapid technological disruption, rationally charting AI's future trajectory is crucial for upholding human agency and flourishing as an intelligent civilization. This paper provides a timely analysis to inform the critical decisions ahead. The paper will discuss the historical background of AI, early approaches, the rise of machine learning, the deep learning revolution, the introduction of large learning models, challenges and ethical considerations, and lastly future directions. Each of these components combined will provide a comprehensive scope of AI.

Historical Background

The foundations of artificial intelligence (AI) can be traced back to the pioneering work of Alan Turing in the 1930s and 1940s. Turing's seminal 1950 paper "Computing Machinery and Intelligence" proposed the idea of the "Imitation Game", later dubbed the Turing Test, as a way to evaluate machine intelligence (Russell and Norvig, 2010). This sparked significant interest in the possibility of creating intelligent machines capable of human-like reasoning and problem-solving. Turing's passion for artificial intelligence was the start of interest in this topic,

which has only increased over time. His research was not mainstream then, yet if it occurred in today's day and age it would have been.

A major milestone was the Dartmouth Conference in 1956, organized by John McCarthy, which is widely regarded as the birthplace of AI as a field of study (McCorduck, 2004).

Attendees from various disciplines proposed different approaches to realizing artificial intelligence, laying the foundations for subsequent developments. Early symbolic approaches, championed by researchers like Allen Newell and Herbert Simon, aimed to hard-code intelligence into systems through logical rules and knowledge representation (Nilsson, 2009).

Allan Newell, Cliff Shaw, and Herbert Simon created The Logic Theorist, which was “designed to mimic the problem solving skills of a human and was funded by the Research and Development (RAND) Corporation. This was presented at the Dartmouth Conference (Anyoha, 2017). Having Dartmouth as the location for a conference such as this provided the field of research with prestige offered by very few institutions.

The 1960s and 70s saw progress in areas like the General Problem Solver and expert systems like DENDRAL for molecular structure analysis (Lindsay et al., 1993). However, the symbolic AI paradigm faced challenges with scaling and representing real-world complexities. This led to the exploration of alternative connectionist approaches inspired by biological neural networks throughout the decade of the 1980s, spurred by the parallel distributed processing framework (McClelland et al., 1986).

The 1990s and 2000s witnessed a resurgence of symbolic techniques combined with statistical learning methods and increasing computational power, enabling major AI breakthroughs (LeCun et al., 2015). Deep learning approaches using large neural networks achieved remarkable performance in areas like vision, speech, and language processing. The

2010s saw transformative advances like ImageNet and natural language models like GPT-3, accelerating AI's real-world adoption across domains (Vaswani et al., 2017; Brown et al., 2020). These models helped push artificial intelligence into the mainstream and made it accessible to large numbers of people.

Early Approaches to AI

The early days of AI were marked by symbolic and logic-based approaches aimed at codifying human knowledge and reasoning into computer programs. Symbolic AI, pioneered by researchers like John McCarthy and Allen Newell, sought to represent knowledge using formal logic, rules, and symbolic manipulation (Russell and Norvig, 2010). Expert systems like MYCIN and DENDRAL were developed to capture human expertise in specific domains, such as medical diagnosis and molecular structure analysis (Buchanan and Shortliffe, 1984; Lindsay et al., 1993). These systems relied on extensive knowledge bases and inference engines to conclude from given facts.

While symbolic AI demonstrated early successes in well-defined problem spaces, it faced significant challenges in handling real-world complexities, uncertainty, and scaling to broader domains. Learning how to overcome these challenges was necessary to bring AI to the next step. Knowledge engineering – the process of manually encoding human expertise into rules – proved labor-intensive and brittle. Additionally, these approaches struggled with perceptual tasks like vision and speech recognition, which require learning from data patterns rather than manual knowledge codification (LeCun et al., 2015). A key limitation was the inability to automatically acquire knowledge and reason about noisy, uncertain environments in the way humans do intuitively. This motivated researchers to explore alternative paradigms like connectionism that

could learn from data **in a more flexible, adaptive manner**. A key pioneering symbolic system was the Logic Theorist in 1956, described as "a program designed to mimic the **problem-solving** skills of a human." However, it remained restricted to formal logical reasoning problems.

Rise of Machine Learning

As the limitations of symbolic AI became apparent, the field witnessed a resurgence of machine learning approaches inspired by biological learning systems and statistical modeling. Key concepts like supervised, unsupervised, and reinforcement learning emerged, enabling systems **to automatically learn patterns and make predictions from data** (Bishop, 2006). **Neural networks**, initially proposed in the 1940s, gained traction as a powerful machine learning framework for modeling complex non-linear functions (McClelland et al., 1986).

Notable early achievements in machine learning included the backpropagation algorithm for training neural networks (Rumelhart et al., 1986), kernel methods like support vector machines (Cortes and Vapnik, 1995), and probabilistic graphical models (Pearl, 1988). These techniques found applications in **areas like** handwriting recognition, speech processing, and recommender systems. However, their performance was often limited by the availability of training data and computational resources at the time. A breakthrough was realizing that these methods could be scaled up by leveraging increased data, model sizes, and compute power - paving the way for the subsequent deep learning revolution. The core principles of letting models learn **their statistical** representations from data were established during this period.

Deep Learning Revolution

The true potential of machine learning, particularly neural networks, was unlocked by the deep learning revolution of the 2010s. As technology progressed, achievement in the field was made easier. Enabled by breakthroughs in hardware (GPUs, TPUs) and software frameworks, researchers could train larger and more complex neural network architectures on massive datasets (LeCun et al., 2015). Techniques like convolutional neural networks (CNNs) for vision tasks (Krizhevsky et al., 2012) and long short-term memory (LSTM) networks for sequence modeling (Hochreiter and Schmidhuber, 1997) propelled deep learning to new heights.

Key advancements included the development of efficient training algorithms (e.g., batch normalization, residual connections), novel architectures (e.g., transformers, generative adversarial networks), and transfer learning approaches (Howard and Ruder, 2018). Deep learning systems achieved unprecedented performance on various AI benchmarks, surpassing human-level capabilities in tasks like image recognition, speech recognition, and game-playing (e.g., AlphaGo defeating world champions) (Silver et al., 2016). This enabled tangible real-world applications spanning computer vision, natural language processing, healthcare, gaming, and more. The ability to learn rich representations from raw data obviated the need for laborious feature engineering. As data and computing resources grew, deep learning models could be scaled to deal with more complex, open-ended reasoning tasks. This led to a proliferation of real-world AI applications leveraging the ability to learn rich representations directly from raw data without manual feature engineering.

Introduction of Large Language Models

A notable development in deep learning has been the emergence of large language models trained on vast textual data using self-supervised learning objectives. Models like GPT-3

(Brown et al., 2020) and PaLM (Chowdhery et al., 2022) have demonstrated remarkable capabilities in understanding and generating human-like text across a wide range of domains and tasks. The self-supervised learning objectives have transformed the way that AI is growing.

These models leverage the transformer architecture (Vaswani et al., 2017) and self-attention mechanisms to capture long-range dependencies and contextualized representations. By pretraining on massive corpora using objectives like masked language modeling, they acquire broad world knowledge and linguistic understanding, enabling few-shot learning and generalization to downstream tasks with minimal fine-tuning (Brown et al., 2020). The scale and multi-task scope of these models, combined with emergent abilities like few-shot learning, opened up novel AI capabilities without the need for arduous per-task engineering.

While still imperfect, large language models have opened up new frontiers in natural language processing, enabling applications like conversational AI assistants, creative writing, code generation, and question-answering. However, they also raise concerns about biases, hallucinations, and the potential for misuse (Bender et al., 2021). Their scale and lack of grounding open up risks like generating deceptive, unsafe, or misaligned outputs without careful controls. Since AI has only become mainstream recently, there is little framework to govern it and ensure that it is safely utilized.

Challenges and Ethical Considerations

Despite remarkable progress, AI still faces significant challenges and limitations, which should not be discounted. Issues like data scarcity, distribution shifts, adversarial examples, and lack of robust generalization hinder the real-world deployment of AI systems (Amodei et al., 2016). As of yet, no solution has been figured out. Furthermore, AI models can exhibit biases

and fairness issues, reflecting societal biases **present** in their training data (Bolukbasi et al., 2016).

As AI systems become more capable and integrated into critical decision-making processes, ethical considerations around transparency, accountability, and value alignment become paramount (Amodei et al., 2016). Concerns around job displacement, privacy violations, and the potential for AI misuse or an uncontrolled intelligence explosion must be carefully navigated (Bostrom, 2014). There is also the risk of modern AI systems operating as **inscrutable** "black boxes", making it challenging to understand how they arrive at outputs and decisions. This lack of interpretability poses challenges for high-stakes decision processes.

Responsible AI development necessitates **a major focus** on safety, robustness, and aligning AI systems with human values from the outset. Multistakeholder collaborations between researchers, policymakers, and domain experts are crucial to **establishing** ethical guidelines, governance frameworks, and regulatory mechanisms for AI (Dafoe, 2018).

Future Directions

The field of AI is rapidly evolving, with numerous exciting frontiers on the horizon. Continued advancements in deep learning architectures, unsupervised learning, multi-**model** AI, and integrating symbolic reasoning with neural networks hold promise for more robust and generalizable AI systems (Marcus, 2020). Emerging paradigms like neuro-symbolic AI, differentiable programming, and meta-learning could lead to more interpretable, adaptable, and data-efficient AI (Garcez et al., 2015; Bengio et al., 2021). **As these are further developed, they will help lead to safer use of AI.**

Additionally, the development of protected exploration techniques, inverse reinforcement learning, and value learning could pave the way for advanced AI systems that can learn and reason about human preferences and values (Hadfield-Menell et al., 2017). Interdisciplinary research at the intersection of AI, neuroscience, cognitive science, and philosophy may yield insights into the nature of intelligence and consciousness (Dehaene et al., 2017). Since AI is gaining popularity, research will likely only increase further in the coming years, allowing for a significantly better understanding of the technology.

However, significant challenges remain, including the lack of a unifying theoretical foundation for AI, the need for more efficient and scalable learning algorithms, and the ethical considerations surrounding the development and deployment of increasingly capable AI systems. Making continued progress while upholding key principles like safety, robustness, interpretability, and value alignment will require sustained multi-disciplinary collaboration bridging domains. Time will determine the rate at which this progress occurs, allowing users to be safer on the different AI platforms.

Conclusion

In summary, the field of AI has undergone a remarkable transformation since Alan Turing's foundational work, fueled by theoretical advances, increases in computational power, and the availability of vast amounts of data. From the early days of symbolic AI and expert systems, the field has witnessed the rise of machine learning, the deep learning revolution, and the recent introduction of powerful large language models. These milestones are key moments in the evolution of AI, demonstrating how much its capabilities have grown.

These advancements have enabled AI systems to match and exceed human capabilities across various domains, from perception and decision-making to natural language understanding and generation. The significant progress in AI has democratized its accessibility, enabling large quantities of people to benefit from its uses. However, the path to more general and robust AI systems remains challenging, requiring a continued focus on safety, interpretability, and value alignment. Ensuring that AI systems are safe is incredibly important to its integration into society properly.

As AI continues to advance, it will profoundly impact various aspects of society, from healthcare and education to transportation and communication. With AI being integrated into different fields, there are many ways that it can help improve the way that they operate. The development and deployment of AI technologies must be guided by ethical principles, regulatory frameworks, and a multistakeholder approach to ensure that AI remains a positive force for humanity.

The future of AI holds both immense promise and significant challenges, but the field's rapid progress over the past few decades instills confidence in the ability to navigate these complexities responsibly. Proactive measures are vital. Determining how to utilize AI safely is paramount to the technology's continued use in society. Continued research, innovation, and multidisciplinary collaboration will be essential in shaping the trajectory of AI to serve the greater good of society and humanity as a whole. Utilizing AI's full potential will help with societal development.

Worked cited

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané.

2016. "Concrete Problems in AI Safety." arXiv preprint arXiv:1606.06565.

Anyoha, Rockwell. "The History of Artificial Intelligence." *Science in the News*, August 28,

2017. <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, 610–623.

Bengio, Yoshua, Andrea Lodi, and Antoine Prouvost. 2021. "Machine Learning for

Combinatorial Optimization: A Methodological Tour d'Horizon." *European Journal of*

Operational Research 290 (2): 405–421.

Bishop, Christopher M. 2006. Pattern Recognition and Machine Learning. New York: Springer.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai.

2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *Advances in Neural Information Processing Systems* 29: 4349–4357.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla

Dhariwal, Arvind Neelakantan et al. 2020. "Language Models Are Few-Shot Learners."
arXiv preprint arXiv:2005.14165.

Buchanan, Bruce G., and Edward H. Shortliffe, eds. 1984. *Rule-Based Expert Systems: The*

MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA:

Addison-Wesley.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam

Roberts, Paul Barham et al. 2022. "PaLM: Scaling Language Modeling with Pathways."

arXiv preprint arXiv:2204.02311.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20

(3): 273–297.

Dafoe, Allan. 2018. "AI Governance: A Research Agenda." *Governance of AI Program, Future*

of Humanity Institute, University of Oxford: 1–35.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider. 2017. "What Is Consciousness, and Could

Machines Have It?" *Science* 358 (6362): 486–492.

Garcez, Artur d'Avila, Tarek R. Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas

Icard, Kai-Uwe Kühnberger, Luis C. Lamb, Risto Miikkulainen, and Daniel L. Silver.

2015. "Neural-Symbolic Learning and Reasoning: Contributions and Challenges." In

2015 AAAI Spring Symposium Series.

Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. 2017.

"Inverse Reward Design." arXiv preprint arXiv:1711.02827.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural*

Computation 9 (8): 1735–1780.

Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-Tuning for Text

Classification." arXiv preprint arXiv:1801.06146.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet

Classification with Deep Convolutional Neural Networks." *Advances in Neural*

Information Processing Systems 25: 1097–1105.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444.

Lindsay, Robert K., Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. 1993. "DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation." *Artificial Intelligence* 61 (2): 209–261.

Marcus, Gary. 2020. "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence." arXiv preprint arXiv:2002.06177.

McClelland, James L., David E. Rumelhart, and PDP Research Group. 1986. *Parallel Distributed Processing. Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*. Cambridge, MA: MIT Press.

McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 2nd ed. Boca Raton, FL: CRC Press.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York: Cambridge University Press.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–536.

Russell, Stuart, and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den

Driessche, Julian Schrittwieser et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–489.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,

Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv preprint arXiv:1706.03762.