

The Voice of the Higher Education Technology Community

EDUCAUSE
REVIEW

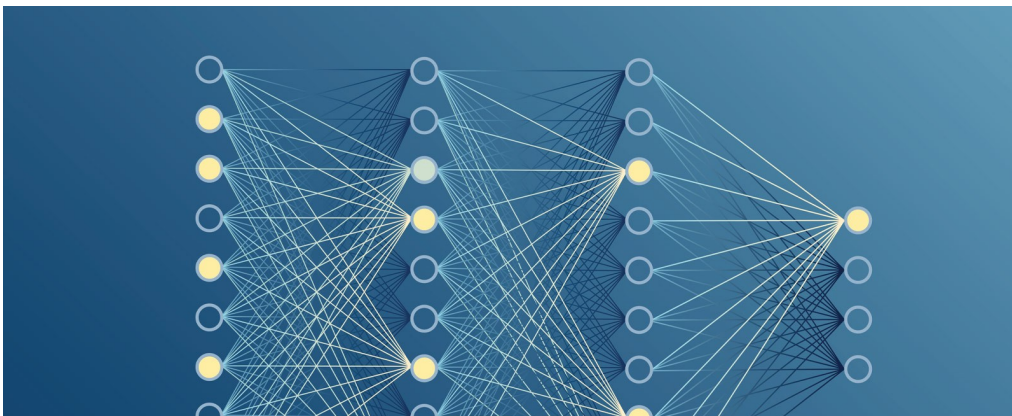
A Generative AI Primer

Brian Basgen Tuesday, August 15, 2023**Emerging Technologies and Trends** ★ Editors' Pick

20 min read



Understanding the current state of technology requires understanding its origins. This reading list provides sources relevant to the form of generative AI that led to natural language processing (NLP) models such as ChatGPT.





Credit: Qpt / Shutterstock.com © 2023

Generative artificial intelligence (AI) is in a renaissance amid a profusion of new discoveries and a breathless frenzy to keep up with emergent developments. Yet understanding the current state of technology requires understanding its origins. With the state of AI science changing quickly, we should first take a breath and establish proper footings. To help, this article provides a reading list relevant to the form of generative AI that led to natural language processing (NLP) models such as ChatGPT.

Artificial neural networks¹ have been around for more than sixty years. We can start in 1958, when the *perceptron* was initially implemented on a computer.² For the first time, an algorithm used computing power to ingest labeled sets of data and classify them into categories. Nearly thirty years later, in 1986, *backpropagation* was introduced as an algorithm for training a neural network to learn from its mistakes.³ This advancement enabled multilayer perceptron networks to learn in nonlinear ways. In 2017, the *Transformer* was created as a neural architecture that improves on the memory limitations of past recurrent models.⁴ This efficiency gain enables a network to better handle the context and relevance of the information it has been provided. All three of these technologies remain relevant and used today.

There is a deep historical record about practical developments

in neural networks,⁵ and the theoretical work supporting them includes centuries of rigorous discoveries in math and biology. Interestingly, despite this rich context of empirical and conceptual work, neural networks have not received much popular attention. AI developments have broken into the collective consciousness only occasionally—for instance, when the IBM supercomputer Deep Blue beat world champion Garry Kasparov in a chess match in 1997 or, more recently, when Google DeepMind's neural network AlphaGo defeated Go grandmaster Lee Sedol in 2016. Instead of AI developments, computer and data networks have commanded most of the attention of technologists in the past few decades. While AI has been relevant and useful in our lives for at least the past ten years, the barrier to entry has been high, and its applicability to higher education has not been obvious. The time has come, however, for IT professionals to focus on AI in its various forms, not because it will solve all problems but because it will transform how organizations operate.

Neural Network Basics

Since this is a reading list, we'll start off with a video! How could we do such a thing? Because at this stage, only a human author can make such a contradictory statement. If we must condense decades of knowledge about neural networks, then it is necessary to focus on the fundamentals. Video is a helpful tool at the outset because neural networks are fundamentally visual. This is partly because neural networks are based on visual math (e.g., calculus, statistics, and linear algebra) and also because the multiple layers of networks are inherently multidimensional concepts.

Produced in 2018 by 3Blue1Brown, the video ["But What Is a Neural Network? | Chapter 1, Deep Learning"](#) occupies an important moment in time: before Transformers began to accelerate NLP models. It presents a classic challenge in AI development: how to handle recognition of handwritten numbers. This video articulates the fundamental components of modern neural networks including the concept of *supervised learning*. It also provides an example of the conceptual problem presented by *hidden layers*.

Now that we've seen a general representation of how a neural network functions, it is time to expand our horizon: there are many types of neural networks, each with particular strengths and weaknesses. [Fjodor van Veen's](#) list, ["The Neural Network Zoo"](#) (The Asimov Institute, September 14, 2016), will at first seem overwhelming, but the goal in reading about these is not to commit this information to memory. Instead, consider each network permutation as being illustrative of the many possible tools that can be used to solve different problems.⁶

The first category of networks that we need to understand in greater detail consists of *recurrent neural networks* (RNNs), developed in the 1980s. RNNs allow for nodes to interact with each other, enabling the network to handle sequential data (e.g., words in a sentence must be in sequence to be understood). The interaction of nodes within the hidden layers enables the network, with sufficient training and fine tuning, to predict future steps in a data sequence. [Eniola Alese's](#) article ["Understanding RNNs Using the Game of Chinese Whispers"](#) (*Medium*, April 13, 2018) is a helpful introduction to the subject.

In "[And of course, LSTM — Part I: The ABC's of the LSTM](#)" [↗] (*Medium*, June 8, 2018), Alese introduces some of the challenges with RNNs—namely that they have memory problems. In 1997, the *long short-term memory* (LSTM) neural network was developed, solving this problem.⁷ LSTM introduced a memory state to enable the network to retain relevant information in order to understand context, while discarding information that is no longer needed.

While LSTM networks can handle much larger inputs, they are slow to train, and they are less efficient running on modern GPUs that are designed for parallel processing. Despite these limitations, LSTMs continue to be relevant and useful for some tasks (as are RNNs). Two decades later, Transformers were created to address some of these limitations. This technology has largely driven the current revolution in NLP. Transformers have some powerful advantages, including the ability to better understand the context of input data due to a self-attention mechanism that is more precise about word dependency and relevance. [Jay Alammar](#) [↗] provides a high-level conceptual overview in "[The Illustrated Transformer](#)" [↗] (June 27, 2018), whereas Peter Bloem gives a nice introduction from a programmer's perspective in "[Transformers from Scratch](#)" [↗] (August 18, 2019).

Definitions

Note: The definitions listed here are for concepts mentioned in the article. They do not cover all aspects of neural networks.

Fundamental Components

- **Hidden Layer:** A generic term for all layers between the input and output layer. Different configurations of layers and neurons result in different types of neural networks.
- **Neuron:** A multi-layer neural network node that receives multiple inputs with distinct parameters, which are processed through an activation function to produce various types of output.
- **Parameters:** The weights and biases associated with each neuron in a network. These constitute the primary means to tune network functionality and behavior through training.
- **Perceptron:** A single-layer neural network node that receives information (input) and, with the help of an activation function and parameters, produces a binary output.

Neural Network Types

- **Artificial neural network:** A type of software that can independently learn at the expense of reduced transparency. In traditional software, a programmer defines fixed parameters and designs a structure based on logic. Neural networks, like other machine learning algorithms, are not deterministic in this way. Instead, they are probabilistic: they can handle novel tasks without explicitly being designed to do so. Training determines how the software operates, which means that a given neural network can behave very differently based on how it is trained.
- **Long Short-Term Memory (LSTM):** A general improvement to the RNN design. Nodes can retain relevant, specific information from previous nodes'

outputs. This is made possible with a dedicated memory mechanism that is discerning (based on training) about what information should be retained, discarded, or output at a given point in time. These networks are able to handle complex sequential data (e.g., several paragraphs of text).

- **Recurrent Neural Network (RNN):** A network design in which nodes within the hidden layer receive partial information about what was processed by previous nodes. This allows for a type of short-term memory wherein a given node may have condensed information from a previous node's outputs. When combined with new input, this enables the network to make predictions about sequential data.
- **Transformer:** A neural network design that ingests all the input it receives simultaneously, analyzes that information to determine which parts are the most relevant (self-attention), and then uses a basic feed-forward network to determine output. The problems of limited memory caused by handling input in a rigid sequential order (e.g., LSTMs and RNNs) are solved by a holistic understanding that enables the transformer to handle highly complex data types.

Neural Network Training

- **Backpropagation:** Developed for supervised learning, an algorithm that computes gradients (essentially efficiencies) in accord with a loss function and propagates these calculations back through all network layers, adjusting parameters in order to improve network accuracy.
- **Few-shot learning:** A variation of supervised learning in which very little labeled data is provided to the neural

network. A sufficiently sophisticated neural network can interpret a well-curated limited dataset and accurately extrapolate future matches for data and labels it has not seen in training.

- **Reinforcement Learning from Human Feedback (RLHF):** A training technique that can be used independently or in combination with other training techniques. The user helps train the behavior of the neural network by rewarding some responses over others, thus influencing the way the network modifies its own weights and biases.
- **Supervised learning:** An ideal training technique for neural networks that are designed for well-defined datasets. Engineers will curate a set of data with labels, which in effect is like providing the questions and answers for a test. When the neural network processes this data, it adjusts its parameters based on its own results. If it succeeds in correctly labeling the data itself, certain connections between nodes will strengthen; if it fails, those connections will weaken.
- **Unsupervised learning:** A quick way to train a network on things like grammar and knowledge without explicit labeling. Engineers curate a set of data and ensure that it is relevant and clean, but no labels are provided. The neural network produces value in this scenario when it is able to successfully identify meaningful patterns in the data. A loss function is used but, in this case, for optimization of the network rather than accuracy per se.

The Building Blocks of ChatGPT

With some history established, we can now review important articles published in the past few years. To start, note that

OpenAI (the creator of ChatGPT) was a nonprofit organization founded in 2015 by well-established industry leaders with the intent to freely collaborate on AI. In June 2018, OpenAI published **"Improving Language Understanding with Unsupervised Learning,"**⁸ which makes the case for *unsupervised learning* using Transformers.⁸ *Unsupervised learning* is the process of providing data to a neural network without telling the network how to understand the data (e.g., giving the network images of animals without labeling them) and configuring the network so that it can figure out, on its own, what is needed. In the case of GPT-1, 7,000 books were fed into the model to train it to accurately predict future words.

Eight months later, OpenAI announced GPT-2. This version demonstrated capabilities that both exceeded expectations and demonstrated emergent capabilities: ways of operating that the developers had not anticipated. This led OpenAI to do something that didn't garner a lot of attention outside of AI researchers: it restricted the release of GPT-2 due to fears of malicious use (e.g., generating spam, producing fake product reviews). In May 2019, a limited version was released, followed in August by a model at something like half its capability (measured in parameters), and in November 2019 the full GPT-2 model was released. Once again using very accessible language, OpenAI describes GPT-2 in **"Better Language Models and Their Implications,"**⁹ which includes a section on policy implications and thus begins to grapple with the societal impact of this technology.⁹

Six months later, in June 2020, OpenAI released GPT-3, without any staged release or restrictions to capabilities. Instead, OpenAI decided to create a for-profit arm and

refrained from releasing the source code to GPT-3. The key breakthrough in GPT-3 is that after increasing the size and configuration of the model, OpenAI discovered that the model was now capable of *few-shot learning*: performing supervised learning with very limited data, enabling rapid adaptation. Thus, while the previously discovered practice of unsupervised learning continued to be followed in GPT-3, the supervised learning portion of training was dramatically improved. Unfortunately, OpenAI did not provide the same high-level summary as it had for GPT-1 and GPT-2, but it did make up for this change of tack with a very thorough technical paper: Brown et al., **"Language Models Are Few-Shot Learners"** [!\[\]\(3dfb8d66e81160ad61421a3452093d1b_img.jpg\)](#) (July 22, 2020). While this paper requires some knowledge in the field, it explains the relevance of this technological advancement.

In March 2022, OpenAI released GPT-3.5, with GPT-4 following just a year later, in March 2023. The chatbot product known simply as ChatGPT—launched on November 30, 2022—for the first time provided an easy, web based interface to the GPT-3.5 model. Something that is notable with these releases (particularly starting with the November 2022 reconfiguration of GPT-3.5) is the dramatic increase in emergent capabilities. This time, OpenAI did not publish a high-level overview for ChatGPT. Instead, the web page **"Introducing ChatGPT"** [!\[\]\(99f58673407353e96a019fbca558fd72_img.jpg\)](#) simply describes what the developers saw as some neat but essentially "iterative" changes.¹⁰

In February 2023, Stephen Wolfram, a pioneer in the development and application of computational thinking, wrote a very thorough blog post: **"What Is ChatGPT Doing . . . and**

Why Does It Work?" ¹⁰ Wolfram summarizes how ChatGPT works. It is a long read with some math, but it is an outstanding resource.

Meanwhile, AI researchers became increasingly excited and concerned about what GPT was capable of. They started to notice that entirely new capabilities were possible. Wei et al. described this phenomenon in **"Emergent Abilities of Large Language Models"** ¹¹ (*Transactions on Machine Learning Research*, August 2022). The co-authors noted societal risks, including behavioral and sociological concerns.

NLP Model Ethics

NLP models like ChatGPT raise a variety of ethical questions. While OpenAI has operated with intentionality with respect to ethical concerns, monetary and corporate interests likely mean that the company's ethics are not guaranteed. Considering the exploding growth of alternative generative AI models, however, many of these products will face much less scrutiny and may have far more significant ethical lapses. But before we can discuss ethics particular to AI, we should consider an ethical framework.¹¹ Hegel's theory of responsibility is a helpful mooring: social context is essential for interpreting ethics, and for any given action, it is also important to consider responsibility for the consequence of this action.¹² Further, knowledge of consequences is imperative for interpreting responsibility.

Starting with the ethical question of bias,¹³ OpenAI has put in significant work to train the GPT model to be mindful of social context and to manage for bias. This is quite unlike Microsoft's

infamous 2016 chatbot Tay, where the problem of AI bias took on an extreme form. The relatively positive results of GPT demonstrate the importance of supervised training. A pressing concern is that competing models have not been trained as intentionally; in addition, "desktop" generative AI models, which are accessible to anyone, could become unethical by design or accident. In **"Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP"** ¹³ (*ACL Anthology*, July 2020), Blodgett et al. provide an excellent critique of bias research in NLP models and make a compelling case for why much more work is needed to improve bias recognition and training in NLP. In **"On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"** ¹⁴ (*FAccT '21 Proceedings*, March 2021), Bender et al. address ethical concerns about increasing NLP models to sizes that may amplify bias and create adverse environmental impacts.

Misinformation is another significant concern. Generative AI models, including GPT, "hallucinate" or confabulate by making up information. This is particularly jarring because the confabulations can appear authoritative: made-up sources sometimes reference real authors but fake titles and even specific page numbers.¹⁴ Such faults can be addressed through supervised training: try asking an NLP model about the company that produced it, and the results will be very consistent. We have also seen GPT's dramatic reduction in confabulation errors as new releases come out, thanks in part to *reinforcement learning from human feedback* (RLHF), in which users rank responses in terms of accuracy. Meanwhile, a product like Bing Chat has attempted to address this issue with a "creativity" setting with the intended goal of confabulation. This is an ethical position that requires nuance,

since it posits that when an NLP makes up information, it is being creative rather than misleading. In **"Training Language Models to Follow Instructions with Human Feedback"** [↗] (March 2022), Ouyang et al. make a compelling case for using RLHF in NLP.

A similar consideration is the impact that this technology has on society. We continue to struggle with the ethical impacts of social media, a technology that is also driven by AI. Since this different form of AI has been used to disseminate fake information, it is apparent that NLP models can be used in a similar manner. Social media AI is, ironically, designed around controlling our attention. One of the consequences of controlling attention is algorithms that promote extreme viewpoints to keep users engaged. This, in turn, leads some people to create socially siloed cultures. When NLP bots are pervasive and anonymized across platforms, how will that change how we interact with each other? What does it mean for our understanding of reality when so much of our society is digital and when the digital output of AI is indistinguishable from the digital output of a human? Will NLP models take a step further than social media and create a distinct AI culture? Hunt Allcott and Matthew Gentzkow provide a thorough account of the role of social media in influencing different kinds of social behaviors in their article **"Social Media and Fake News in the 2016 Election"** [↗] (*Journal of Economic Perspectives*, Spring 2017).

Privacy and security are complicated ethical concerns for NLP, since we understand very little about the possible consequences of emergent capabilities. Spam and phishing are obvious concerns: grammar mistakes will no longer be a


telltale sign of fake solicitations. There are also immediate concerns about the data directly shared with the model (and breaches have already occurred). Yet the ethical questions run far deeper. To what degree can exposure to limited information about a person allow for excellent impersonation of that person, rendering so much of our security infrastructure invalid? To what extent can a model learn about an individual and accurately predict how that person thinks and acts in order to manipulate someone to take an action? How much information will be necessary to determine someone's relationships, habits, and preferences in ways that surveillance states could leverage? If the business model of social media AI is advertising (and thus, AI is influencing people to buy products), what business models will emerge for NLP? How will the mechanism of autonomy on the internet, sometimes used by humans in problematic ways, allow an effective mechanism of responsibility for AI? Indeed, autonomy and privacy are likely on the verge of reformation. Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*¹⁴ (February 2018), provides a comprehensive overview of various security problems posed by AI. While the report is long, it is essential reading and provides a framework that will hopefully be updated in the future.

The crux of AI ethics is accountability and transparency.¹⁵ A key component in neural network functionality is the hidden layers that lie beyond the direct control of programmers. When emergent behavior happens, the practical result is that a black box effect has been extended to the output layer: unanticipated behavior or capability has been realized. There is an important ethical distinction between not knowing how

something is done and not knowing what the network is able to do. For every novel permutation of different outputs, we have no mechanism to evaluate the ethical implications of the consequences that may follow. Whether we follow Hegel or other ethical philosophers, understanding how these models produce specific results and understanding the full spectrum of what they are capable of producing is essential.¹⁶

Technologists looking to use NLP models within their institutions need to establish an ethical framework that considers those concerns within their realm of agency and control. One step toward transparency is attribution: how and when do we attribute work to generative AI? Looking up a topic up on the internet, getting ideas, and writing about it is very different from copying sections from an article without attribution. Similarly, direct copying from content produced by generative AI tools requires disclosure, but using that content for ideation does not. Consider also the need for accountability. Blindly trusting in generative AI output is unwise and often unethical. AI output cannot be cited because it is not referenceable. Validating AI output by consulting reliable sources to arrive at a sound conclusion is surely reasonable.¹⁷

Ethical Guidelines on the Use of Artificial Intelligence (AI) and Data in Teaching and Learning for Educators  (2022), published by the European Commission, is an excellent guide for the thoughtful use of AI in education.

Since I started this reading list with a video, how could I pass up the opportunity to end the same way? In the hour-long video "**The A.I. Dilemma**"  (March 9, 2023), Tristan Harris and Aza Raskin, two cofounders of the Center for Humane Technology, excellently summarize the dangers currently

posed by the AI field generally. They discuss what the future adoption of these tools may look like based on what we know now about the current capabilities of AI technology.

In Summary

The history of AI is entwined with the history of computing. While the importance of the internet is now widely accepted, the internet is a new technology by comparison with AI. The societal relevance of the internet languished for more than two decades after its creation, yet the incubation period for AI has been even longer.

AI is no longer a specialized or niche field of research with limited applicability. In the past decade, AI has been delivered, in one form or another, to consumers without the AI label (e.g., social media platforms, voice assistants, recommendation systems, photo editing). ChatGPT is an announcement to the world that the field of deep learning—that is, highly capable neural networks—is now immediately accessible to all. Perhaps more importantly, ChatGPT is a demonstration of the effectiveness of one type of AI tool that will encourage all of us to invest in and explore this expansive landscape of possibilities.



As many of us in IT organizations embrace AI tools in the coming years, we need to recognize the consequences of our actions. In time, AI will fundamentally challenge the way we understand reality. AI is a novel collection of technologies without historical parallel. Just as AI will not solve all problems, there is not one simple and ethical way to navigate AI. Instead, a rigorous analysis of ethics and metaphysics and a

careful consideration of how IT organizations deploy AI will be required. IT organizations need to build on the ethical framework of equity and inclusion to ensure that this empowering technology experience of generative AI is transparent and accountable.

Notes


1. For the purposes of this article, *neural networks* throughout will refer to artificial neural networks. ↩
2. See Warren S. McCulloch and Walter Pitts, "**A Logical Calculus of the Ideas Immanent in Nervous Activity**," [!\[\]\(f15d3c54be60b4fd0ce1da9fb3f67256_img.jpg\) *Bulletin of Mathematical Biophysics* 5 \(1943\)](#), as the historical introduction to this concept. While this is principally a theoretical math paper and there are better ways to understand how a perceptron functions, the authors understood they were writing to a much wider audience, addressing philosophers and psychologists. Consider the statement about neural networks: "Thus our knowledge of the world, including ourselves, is incomplete as to space and indefinite as to time. This ignorance, implicit in all our brains, is the counterpart of the abstraction which renders our knowledge useful. The role of brains in determining the epistemic relations of our theories to our observations and of these to the facts is all too clear, for it is apparent that every idea and every sensation is realized by activity within that net, and by no such activity are the actual afferents [nerve fibers] fully determined." ↩
3. See D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "**Learning Internal Representations by Error Propagation**," [!\[\]\(7bf135d42c40a6430c927b2fd03d7659_img.jpg\)](#) in D. E. Rumelhart and J. L.

McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition; Vol. 1: Foundations* (Cambridge, MIT Press, 1986). This paper is technical, with some heavy math sections, but the authors have written it for a wider audience. The paper does assume background knowledge in the field, but it also serves as an important historical marker in that it addresses issues identified in previous work. Most famously, by explaining how multilayer perceptrons can be successfully trained, the paper provides a satisfying rebuke to Marvin Minsky and Seymour Papert's 1969 critiques of neural networks driven by single-layer perceptrons. In this manner, AI research had fully emerged from the first "AI winter" of the 1970s by embracing multilayer networks that could solve novel problems. ↩

4. The seminal 2017 paper "**Attention Is All You Need**"  by Ashish Vaswani et al. gives an introduction to this subject. Despite the catchy title, this paper is quite technical and requires significant working knowledge of deep learning. Note that Google pioneered this Transformer technology and made it open source, along with other key AI technologies. ↩
5. A. Ivakhnenko and V. G. Lapa's book ***Cybernetic Predicting Devices***  (Kiev, 1965) was developed in the Soviet Union and was the genesis of deep learning. For the first time, a mathematical foundation was established for how neural networks could successfully operate with multilayer perceptrons. While this is a technical work, in Chapter 4 the authors recognize the importance of Frank Rosenblatt's 1958 work with implementing perceptrons, and they provide a great

contemporary understanding of the model. Following up on this work was Ivakhnenko's article "**Heuristic Self-Organization in Problems of Engineering Cybernetics**," [!\[\]\(c507f772dba2b921f86777f01218e570_img.jpg\)](#) *Automatica* 6 (1970), in which he introduced an algorithm to automate neural network construction and utilized the theoretical groundwork he had previously established with multilayer perceptrons.



6. A wonderful resource for more information on other types of networks is the blog series on deep learning written by Tim Dettmers, a staff member at NVIDIA, in 2015–2016. The **first blog post** [!\[\]\(aca6fcc8bd95e8255b9ea1b1d08ef300_img.jpg\)](#) focuses on visual AI (e.g., how convolutional artificial neural networks function). The **second blog post** [!\[\]\(0083087c61cec498ac803a4aec5bb1bd_img.jpg\)](#) provides a history of deep learning and discusses some of the math developments. The **third blog post** [!\[\]\(2e94242fda9f31152eb2b29146bfce46_img.jpg\)](#) covers some of the fundamental building blocks of NLP models. The **fourth blog post** [!\[\]\(680c68b4e62fe5ec9774c1168e904fbf_img.jpg\)](#) discusses different network training methods of the time. 
7. For a deeper dive into the subject, read Sepp Hochreiter and Jürgen Schmidhuber, "**Long Short-Term Memory**," [!\[\]\(87f26857125315836dd413b717a8c1ec_img.jpg\)](#) *Neural Computation* 9, no. 8 (1997). While this is a very technical paper, it can be usefully navigated by the uninitiated. The authors do a great job of providing context for the problem, and they explain previous work up to that point. This provides an informative but condensed historical summary that illustrates the various obstacles early networks encountered. Section 4 outlines their core proposal with math but also offers a conceptual explanation of how LSTMs work. Section 6 contains a remarkably prescient set of LSTM limitations and advantages that have remained relevant

for twenty-five years. ↩

8. This high-level summary of their work is quite accessible to nonexperts. For the technical version, see Radford et al., **"Improving Language Understanding by Generative Pre-Training"** [↗](#) [2018]. ↩
9. For the technical version, see Radford et al., **"Language Models Are Unsupervised Multitask Learners"** [↗](#) [2019]. An interesting way to look at this paper is to search for the term *surprising*, as noted by Benj Edwards in **"Why ChatGPT and Bing Chat Are So Good at Making Things Up,"** [↗](#) *Ars Technica*, April 6, 2023. ↩
10. The March 2022 initial release of 3.5 received similarly modest treatment by way of **feature announcement.** [↗](#)
↩
11. Ethical philosophy is absent in many publications on AI ethics, and this often results in shallow ethical discussions. Furthermore, since 2021 several ethicists with robust philosophy backgrounds have been fired by both Google and Microsoft. See Daisuke Wakabayashi and Cade Metz, **"Another Firing among Google's A.I. Brain Trust, and More Discord,"** [↗](#) *New York Times*, May 2, 2022; Zoe Schiffer and Casey Newton, **"Microsoft Lays Off Team That Taught Employees How to Make AI Tools Responsibly,"** [↗](#) *The Verge*, March 13, 2023. ↩
12. Georg Wilhelm Friedrich Hegel, *Elements of the Philosophy of Right* (1820). Hegel's thoughts on imputability are also relevant. Knowledge of what may occur as a result of one's actions is an important discourse when considering emergent behaviors in AI.



13. For a recent historical perspective, see Sofiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018), a groundbreaking book based on years of research about Google's search algorithms and the pervasive bias within them.
14. Oddly, for some famous confabulations that have occurred to date, companies disclaimed responsibility. This is a dubious ethical position because the only substantive function of NLPs is their output. This is like saying the company is responsible only when there is no problem! Since the summer of 2023, all major models now feature disclaimers about this issue.
15. Several companies responsible for generative AI, including OpenAI, made bold statements in the spring of 2023 about the existential risks posed by the technologies they have developed. This moral dissonance is troubling, if not hypocritical. Dr. Sasha Luccioni, a machine-learning research scientist at Hugging Face, adds: "It's also misdirection, attracting public attention to one thing (future risks) so they don't think of another (tangible current risks like bias, legal issues and consent)." (Quoted in Benj Edwards, **"OpenAI Execs Warn of 'Risk of Extinction' from Artificial Intelligence in New Open Letter,"** *Ars Technica*, May 30, 2023.)
16. ***Generating Harms: Generative AI's Impact and Paths Forward*** (Electronic Privacy Information Center, May 2023) is a contemporary overview of threats posed by generative AI. This report provides practical examples of how generative AI can be misused, which is

particularly helpful when considering institutional adoption of these technologies. ↩

17. Interestingly, ChatGPT works within an explicit knowledge cutoff date of September 2021. This is partly a practical matter, since pretraining has to begin and end at some point in time. It is also, however, intentional: OpenAI researchers understood that these models would soon be used to populate the internet with generated content and that training a neural network on content generated by other neural networks leads to model collapse. Similarly, if we end up "validating" the results of ChatGPT through web searches that are driven by other NLP models and lead to AI-generated content, does that really count as validation? ↩
-

Brian Basgen is CIO at Emerson College.

© 2023 Brian Basgen. The text of this work is licensed under a **Creative Commons BY-NC-SA 4.0 International License**. ↗

► **Artificial Intelligence (AI), Infrastructure and Research Technologies**

