LEGISLATIVE OFFICE BUILDING
1020 "N" STREET, SUITE 162
SACRAMENTO, CA 95814
(916) 319-2200

PRINCIPAL CONSULTANT
JULIE SALLEY

COMMITTEE SECRETARY
MIMI HOLTKAMP

Assembly
California Legislature

ASSEMBLY COMMITTEE ON
PRIVACY AND CONSUMER PROTECTION
REBECCA BAUER-KAHAN, CHAIR
ASSEMBLYMEMBER, SIXTEENTH DISTRICT

INFORMATIONAL HEARING
PRIVACY AND CONSUMER PROTECTION COMMITTEE

UNDERSTANDING AI: MYTHS, MAGIC, AND MACHINE LEARNING

February 27, 2024
1:30 p.m.
O Street, Room 1100

BACKGROUND PAPER

I. INTRODUCTION

The development of artificial intelligence (AI) is creating exciting opportunities to grow California's economy and improve the lives of its residents. AI can generate compelling text and convincing images in an instant. It can automate painstaking tasks, identify subtle patterns in large datasets, and make accurate predictions in the face of incomplete information. AI is driving cars in San Francisco, fighting fires in the Central Valley, and recording voiceovers in Hollywood.[1,2,3] Just as the widespread adoption of the internet ushered in an era of commercial dominance thirty years ago, AI could deliver a second technological golden age to California.

But with novel technologies come novel safety concerns. As AI becomes ubiquitous in society, can Californians be confident that their personal information is not captured and sold without consent? Are the automated decision tools that influence their lives trained on unbiased datasets? When government partners with industry to develop AI tools for its workforce, who ultimately owns these models? The widespread adoption of AI has the potential to transform California for the better – but without careful implementation, these changes could also erode the privacy of Californians and widen existing equity gaps.

---

[1] Tripp Mickle, Yiwen Lu and Mike Isaac, "'This Experience May Feel Futuristic': Three Rides in Waymo Robot Taxis," *New York Times.* August 21, 2023, https://www.nytimes.com/2023/08/21/technology/waymo-driverless-cars-san-francisco.html.
[2] Pranav Dixit, "Stopping Wildfires: AlertCalifornia and Cal Fire AI Wildfire Detector," *Time,* October 24, 2023, https://time.com/collection/best-inventions-2023/6327137/alertcalifornia-ai-wildfire-detector/.
[3] Matt O'Brien, "Some video game actors are letting AI clone their voices. They just don't want it to replace them," *AP News,* February 19, 2024, https://apnews.com/article/aigenerated-voice-clones-video-game-actors-replica-studios-sagaftra-517cc248f60a2f5e35f9b239b70f20a7.

The emerging AI industry will be critically important to California's economy in the coming years, but failing to properly scrutinize a novel technology can have lasting consequences.[4] Enacting a comprehensive legislative framework for AI will allow Californians to engage with AI without fear for their personal safety, while also granting the AI industry full regulatory clarity as it innovates and expands.

AI is math, not magic – it can be understood and regulated. The purpose of this informational hearing is to provide a high-level overview of AI: what it is, what it is not, its applications, its benefits, and its risks. California's AI industry could flourish or flee based on the actions of the Legislature this year, and as a result, it is critically important that California adopt a clear, consistent, and technically feasible approach to AI.

II. THE NOT-SO-SECRET HISTORY OF AI

The old world was slow: the swiftest militaries moved at the speed of a horse, the fastest projectiles moved at the speed of an arrow, and society progressed at the speed of the human brain. The 20th century brought societal challenges that, for the first time in history, brains alone could not address. Computers were invented in the 1940s and 1950s to calculate the trajectory of intercontinental ballistic missiles and break encrypted messages. They were everything that brains were not: huge, precise, and inflexible. As a concept, "artificial intelligence" (AI) – the mimicking of human intelligence by artificial systems – has existed since antiquity.[5] As a science, the AI field emerged in the 1950s as an attempt to combine the processing power of computers with the flexibility of the human brain. Alan Turing, a founding father of computer science, was optimistic that AI could eventually match human intelligence:

> I believe that in about fifty years' time it will be possible to programme computers . . . to make them play the imitation game [a test of whether computers can think] so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning . . . I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.[6]

Early AI research was marked by groundbreaking theoretical advances and significant technological challenges. The term "artificial intelligence" was first coined at a Dartmouth conference in 1956, and by the 1970s the theory of AI had been firmly established.[7] Two barriers stood between theory and reality:

---

[4] Alan Z. Rozenshtein, "Interpreting the ambiguities of Section 230," Brookings Institute, October 26, 2023, https://www.brookings.edu/articles/interpreting-the-ambiguities-of-section-230/.

[5] Alex Shashkevich, "Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths," *Stanford News*, February 28, 2019, https://news.stanford.edu/2019/02/28/ancient-myths-reveal-early-fantasies-artificial-life/.

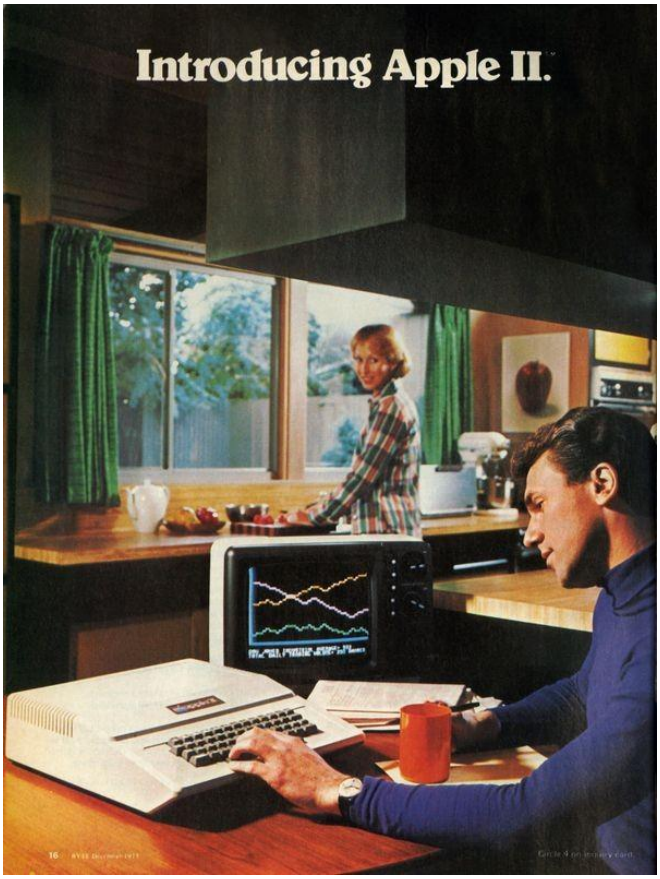
[6] Alan M. Turing, "COMPUTING MACHINERY AND INTELLIGENCE," *Mind*, vol. LIX, no. 236, October, 1950, p. 433-460.

[7] History of Data Science, "Dartmouth Summer Research Project: The Birth of Artificial Intelligence," www.historyofdatascience.com/dartmouth-summer-research-project-the-birth-of-artificial-intelligence/, accessed on February 22, 2024.

1. Computers were underpowered; the hardware that would eventually support advanced AI systems had not yet been invented.

2. High-quality training data was scarce. This data existed in various forms around the world, but there was no single repository that could be easily adapted for training advanced AI.

Both limitations would be overcome in the following decades. 1977 saw the release of the Apple II computer, one of the first commercially successful personal computers (PCs). PCs democratized access to powerful computing resources for the first time. Their widespread adoption ignited a surge in grassroots programming and experimentation, allowing a diverse set of thinkers and tinkerers to contribute to the field of AI. Some of the world's most valuable companies – Apple, Microsoft, and IBM – ascended in this era. PCs steadily became more powerful and affordable, and in the 1980s the internet emerged as a global network capable of connecting people and transmitting information instantaneously. The rise of the internet ushered in the era of data: vast, unstructured, and immensely valuable. The 1990s and 2000s birthed internet powerhouses such as Amazon, Google, and Facebook. These companies cut their teeth on AI-powered search engines and recommendation systems.



Source: https://www.computerhistory.org/revolution/personal-computers/17/300/1049

As PCs and the internet forged the connective tissue of a new global communications network, significant strides were being made in natural language processing – the ability of computers to

interpret, manipulate, and comprehend human languages. In the 1980s speech recognition technologies began to evolve from rudimentary, rule-based systems (*if X is true, then do Y*) into highly accurate AI tools capable of understanding and transcribing human speech. Handwriting analysis transformed from a niche scientific study into a widely-used technology, with applications ranging from personal devices to forensic science. The development of optical character recognition (OCR) made it possible to automatically transfer text from physical documents into digital repositories. Millions of legal documents, health records, and old books were uploaded to the internet in a format that enabled quick searching and editing, where previously these efforts would have required painstaking manual transcription. Combined with twenty years of forum conversations, websites, and various other forms of written content, these digitized documents turned the internet into an unparalleled repository of human knowledge – and the perfect dataset for training an advanced form of AI.

The final technological milestone on the road to the modern era was the rise of the Graphics Processing Unit (GPU) in the 2000s. GPUs were not invented with AI in mind – in fact, they were originally developed by companies such as Nvidia and AMD for the video game industry. While traditional Central Processing Units (CPUs) are designed to handle sequences of tasks (*calculate X, then calculate Y, then calculate Z*), GPUs can process many tasks simultaneously (*calculate X and Y and Z*). In the 2010s massive internet datasets dovetailed with cheap, plentiful computing resources to permit an advanced form of AI: deep learning. Through deep learning, AI could be trained to make complex decisions and perform human-like tasks.

AI had always been integral to the function of the internet. Thirty years earlier, targeted advertising had revolutionized marketing by using AI to analyze consumer behavior and preferences, while AI-powered recommendation systems had driven social media feeds and online shopping. Despite these advances, AI's ubiquity remained largely unappreciated by the general public. This changed in 2022 with the release of ChatGPT 3.5, a generative AI (GenAI) tool developed by the nonprofit OpenAI.[8] ChatGPT could engage in lengthy conversations with its users. It could write poetry, jokes, and computer code. Having been trained on the entirety of the public internet, there was little it did not know – or that it would not pretend to know. A second OpenAI product, DALL-E, could generate intricate images of any subject, in any style, in seconds.[9] These releases had an immediate effect on society. Coalitions of industry leaders and public figures called for moratoriums on the development of more powerful AI, to no effect; OpenAI had let the genie out of the bottle, and it would not be coaxed back in. [10]

III. HOW DOES AI WORK?

AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. This is because AI is not fundamentally different from other computer functions. Its novelty lies in its application: unlike normal computer functions, AI is able to accomplish tasks that are normally performed by humans. Calculation and data storage are normal computer functions. Software that interprets

[8] OpenAI, "Introducing ChatGPT", openai.com/blog/chatgpt, accessed on February 22, 2024.
[9] OpenAI, "DALL-E now available in beta", openai.com/blog/dall-e-now-available-in-beta, accessed on February 22, 2024.
[10] Cade Metz and Gregory Schmidt, "Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'," *New York Times*, March 29, 2023, https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

emotions in human faces is AI.[11] Any bill seeking to broadly regulate AI should understand the industries and applications the policy will impact; these include, but are not limited to:

Pattern recognition

- Drug design
- Medical diagnosis
- Disease surveillance
- Fraud detection
- Sentiment analysis
- Online content moderation
- Nuclear stockpile maintenance
- Environmental monitoring
- Sports analytics

Decision making

- Credit scoring
- Rate setting (insurance, mortgages, etc.)
- Supply chain management
- Retail inventory management
- Traffic management
- Energy management
- Cybersecurity
- Military strategy
- Game theory

Text and image generation

- Chatbots
- Educational materials
- Articles, blogs, and reports
- Coding
- Digital art
- Game development
- Fashion and design
- Advertising and marketing
- Film and entertainment (CGI, voice acting and voiceovers, script writing, etc.)

Natural language processing

- Speech recognition
- Voice-activated assistants (Siri, Alexa, etc.)
- Transcription and stenography
- Language translation
- Content summarization

[11] Tate Ryan-Mosley, "AI isn't great at decoding human emotions. So why are regulators targeting the tech?," *MIT Technology Review,* August 14, 2023, https://www.technologyreview.com/2023/08/14/1077788/ai-decoding-human-emotions-target-for-regulators/.

- Voice generation

Predictive analytics
- Weather forecasting
- Climate modeling
- Market trend analysis
- Energy consumption forecasting
- Agricultural forecasting
- Crime prediction

Recommendation systems
- E-commerce products
- Streaming content suggestions
- Social media feeds
- Online advertising
- News aggregators
- Job matching platforms
- Health and fitness apps
- Retail banking services

Robotics
- Manufacturing
- Autonomous vehicles
- Healthcare and surgery
- Agricultural machinery
- Warehouse management
- Domestic robots (e.g. Roomba)
- Social robots (e.g. hitchBOT)
- Search and rescue operations
- Military and law enforcement surveillance and reconnaissance
- Autonomous weapons
- Defusing and disposing of explosives
- Space exploration

Machine vision
- Quality control (manufacturing, agriculture, retail, etc.)
- Medical imaging
- Retail checkout systems
- Augmented reality and virtual reality
- Document analysis
- Handwriting analysis
- Biometric identification (e.g. facial recognition technology)
- Preservation of art and cultural heritage

Symbolic AI. Consider a standard decision-making task: a game of chess. A major milestone in AI development was reached in 1997, when IBM's Deep Blue computer first beat world

champion Garry Kasparov in a six-game chess match.[12] Deep Blue employed symbolic AI, a form of AI that uses logic and symbols to model the world in order to solve puzzles and prove theorems.[13] This type of AI relies on explicitly programmed instructions and rules to mimic human reasoning. Symbolic AI does not learn from training datasets, and it does not adapt as it interacts with its environment. Symbolic AI has been mostly replaced by machine learning in the modern era.

Source: https://aibusiness.com/ml/25-years-ago-today-how-deep-blue-vs-kasparov-changed-ai-forever

Machine learning. During machine learning, an AI tool is exposed to data and allowed to explore its structure. Unlike symbolic AI, machine learning tools develop their algorithms automatically.[14] The process of exposing a machine learning tool to data is known as "training." The algorithm that a machine learning tool develops during training is known as its "model."

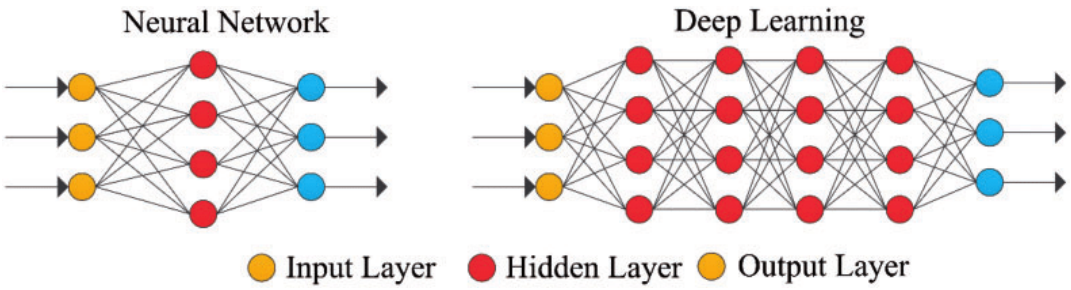
Models that are trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from "generative AI," which are trained on massive datasets in order to produce detailed text and images. When Netflix suggests a TV show to a viewer, the recommendation is produced by predictive AI that has been trained on the viewing habits of Netflix users.[15] When ChatGPT generates text in clear, concise paragraphs, it uses generative AI that has been trained on the written contents of the internet.[16] These systems' architecture and training data differ, but their core concept is the same.

The importance of data. There is a famous saying in computer science: "garbage in, garbage out." The performance of an AI tool is directly impacted by the quality, quantity, and relevance

[12] IBM, "Deep Blue," www.ibm.com/history/deep-blue, accessed on February 22, 2024.
[13] John Markoff, "Innovators of Intelligence Look to Past," *New York Times*, December 15, 2014, https://www.nytimes.com/2014/12/16/science/paul-allen-adds-oomph-to-ai-pursuit.html.
[14] IBM, "What is machine learning?," www.ibm.com/topics/machine-learning, accessed on February 22, 2024.
[15] Netflix, "How Netflix's Recommendations System Works," help.netflix.com/en/node/100639, accessed on February 22, 2024.
[16] OpenAI, "How ChatGPT and Our Language Models Are Developed," help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed, accessed on February 22, 2024.

of the data used to train it.[17] Before training, data is often preprocessed to make it easier for the AI to work with. Rigorously preprocessing a dataset becomes more difficult as the dataset becomes larger, but failing to rigorously preprocess and review the data can lead to meaningless, false, or biased outputs.

The advent of deep learning. The late 2010s witnessed a revolution in deep learning, a subset of machine learning that uses multilayer (deep) neural networks to analyze vast datasets.[18] Neural networks are inspired by the structure of the human brain. They have three components: an input layer that receives information, a hidden layer that processes information, and an output layer that produces the final decision or prediction. Deep neural networks have many hidden layers. They transfer input data from layer to layer, transforming it along the way.



Source: https://journals.sagepub.com/doi/10.1177/0735633118757015

Each layer of a deep neural network contains many nodes, or neurons. These nodes are the fundamental processing units of the network. In most networks, each node is connected to every node in the next layer. The strengths of these connections – the degree to which individual nodes influence their partners in the next layer – are referred to as a model's "weights". These weights provide the basis for learning, as they are continuously adjusted during training to improve a model's performance.

Foundation models. Foundation models are AI tools that have been pre-trained on extensive datasets covering a wide range of knowledge and skills. Once trained, these models serve as a "foundation" that can be fine-tuned for specific tasks. It is useful to think of foundation models from a nature vs. nurture perspective: the "nature" of the models is baked in during the original training process, but the model's ultimate purpose and behavior can be "nurtured" through further training. Whether the tool becomes helpful or harmful depends on the type of data used to nurture it. For example, when a foundation model is trained on patient records and doctors' notes, it can become a tool that helps doctors correctly diagnose patient symptoms.[19] On the other hand, when the same model is nurtured using misogynistic and racist data – data that might appear on a website like 4chan – the tool becomes a hateful alt-right meme generator.[20]

---

[17] Rohit Sehgal, "AI Needs Data More Than Data Needs AI," *Forbes*, October 5, 2023, https://www.forbes.com/sites/forbestechcouncil/2023/10/05/ai-needs-data-more-than-data-needs-ai/.
[18] Amazon Web Services, "What is Deep Learning?," aws.amazon.com/what-is/deep-learning/, accessed on February 22, 2024.
[19] Hanyin Wang et. al, "DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients," *njp digital medicine*, vol. 7, no. 16, 2024.
[20] Andrey Kurenkov, "Lessons from the GPT-4Chan Controversy," *The Gradient*, June 12, 2022, thegradient.pub/gpt-4chan-lessons/.

Open-source vs. closed-source models. Foundation models can be released as open-source or closed-source products by their creators. Open-source models are publically available; researchers and developers can access the models' code and pre-trained weights. This accessibility increases transparency, but it has downsides: when a model's code and weights can be easily accessed they can be easily altered, and open-source models have the potential to be used for nefarious purposes such as generating deepfake pornography and targeted propaganda.

Open source models currently on the market include:

- **LLaMA2:** "Large Language Model Meta AI 2". An open source LLM released by Meta in 2023.[21]
- **Bert:** "Bidirectional Encoder Representations from Transformers". An open source LLM released by Google in 2018.[22]
- **Stable Diffusion:** An open source image generator released by startup Stability AI in 2022. In contrast to DALL-E and Midjourney, who employ GANs to generate images, Stable Diffusion utilizes an architecture known as a diffusion model.[23]

Conversely, closed-source models are opaque with respect to their security features and input/output filters. It is harder for bad actors to generate illicit materials using these models. However, unlike open-source models, closed-source models are not subject to collective oversight because their inner workings cannot be examined by independent authorities. For example, an automated decision tool that produces biased outputs can be harder to detect and correct if the tool is closed-sourced.

Closed-sourced models currently on the market include:

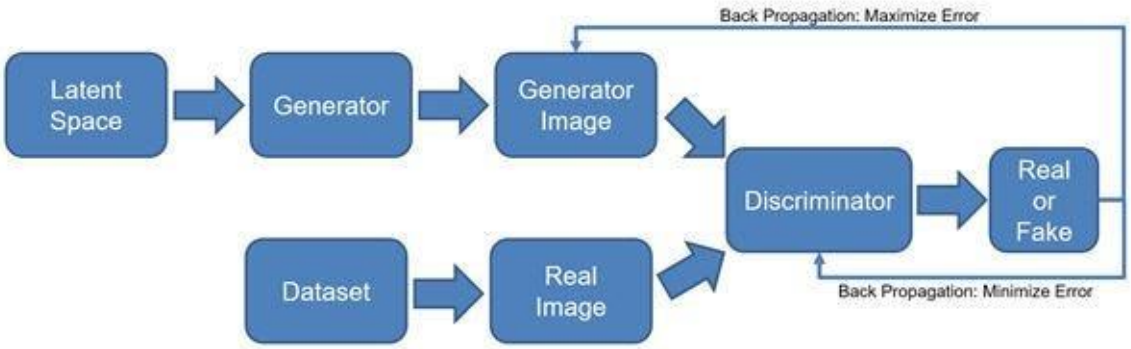
- **GPT-4:** A large multimodal model that accepts both image and text inputs and emits text outputs. Released by OpenAI in 2023.
- **AlphaFold:** Released by Google DeepMind in 2018, AlphaFold predicts protein structures from amino acid sequences.
- **Copilot:** A chatbot released by Microsoft in 2023.
- **IBM Watson:** A question-answering system developed by IBM to compete on Jeopardy in 2011, and released commercially in 2013.
- **Siri:** A digital assistant released by Apple in 2011.

Large language models. Large language models (LLMs) are a type of foundation model that has been specifically designed to understand, generate, and work with human language. These models are trained on vast quantities of text sourced from the internet and historical literature. Over the course of training, LLMs can learn the statistics and semantics of languages contained in their training data.

[21] Meta, "Introducing LLaMA: A foundational, 65-billion-parameter large language model," ai.meta.com/blog/large-language-model-llama-meta-ai/, accessed February 22, 2024.
[22] Google Research, "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing," blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html, accessed February 22, 2024.
[23] Stability AI, "Stable Diffusion Public Release," stability.ai/news/stable-diffusion-public-release, accessed February 22, 2024.

In order to produce meaningful outputs, LLMs must understand not only the sequence and meaning of each word in a dataset, but also their context. LLMs employ a specific architecture known as a "transformer" that preserves relationships between the various parts of an input – for example, the individual words in a sentence – even as the input data is chewed up and processed by the model.

Image generators. Just as LLMs create new text, image generators use deep neural networks to create new images based on the patterns and features they have learned from large datasets of existing images. These networks often employ a specific architecture known as a Generative Adversarial Network (GAN), wherein two networks work together to produce detailed images.[24] GANs can be envisioned as a contest between an artist (the generator) and an art critic (the discriminator). During training, the artist tries to create a fake painting that looks real, while the critic judges the art to determine whether it is fake. Through training both individuals improve, until the images produced by the artist are indistinguishable from the original art. The generator half of the GAN (the artist) can be deployed as an image-generating tool, like DALL-E or Midjourney, while the discriminator half of the network (the critic) can be used in AI detection software.



Source: https://medium.com/@Packt_Pub/inside-the-generative-adversarial-networks-gan-architecture-2435afbd6b3b

IV. INCREDIBLE PROMISE OR UNPRECEDENTED THREAT?

AI is already embedded into most online systems. It is integral to many aspects of modern society, and the advent of GenAI will undoubtedly lead to an even greater number of applications. New industries will emerge to curate training data, design hardware, and deploy AI over the next few years; these industries could prove a huge economic boon for California. However, AI is not an inherently benevolent technology – it is a tool, and it can be used for good or ill. Policymakers will need to design regulatory guardrails to limit harmful uses while allowing for the development and refinement of tools that benefit society.

The following section briefly discusses the various negative aspects of AI. The intent of outlining known harms and negative aspects is not to suggest that the harms outweigh the potential benefits of AI tools; rather, it is to highlight areas that might benefit from targeted legislation.

[24] Amazon Web Services, "What is a GAN?," aws.amazon.com/what-is/gan/, accessed on February 22, 2024.

Bias and discrimination. An AI tool is only as good as its training data, and if the data used to train a model is biased, the model's outputs will be similarly biased. Over the past thirty years, several industries have been forced to contend with this fact as they have tried to introduce automated decision tools into their workflows.

1. Hiring and recruitment tools: It is no secret that people of various races, genders, and cultures are not distributed equally throughout the workforce. An AI trained on historical data to make hiring decisions will be predisposed to maintain the ratios it is trained on; as described by Aditya Malik, the Founder and CEO of Valuematrix.ai:

> Generative AI, for all its grandeur, has the potential to perpetuate latent biases inherited from human creators. A disconcerting echo of historical prejudices may inadvertently seep into the algorithms. Imagine a scenario where previous senior managers, driven by biases of gender, age, faith or race, rejected candidates for misguided reasons. The AI, if not vigilantly curated, might misconstrue these patterns as indicators of incompetence, thus exacerbating the exclusion of qualified candidates from underrepresented backgrounds.[25]

This was notoriously experienced by Amazon, who attempted to automate their hiring practices in the early 2010s. They scrapped this approach in 2015 when they realized that their new system was not rating candidates in a gender neutral way. In fact, their system was excluding women from the pool of acceptable candidates because it had been trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.[26]

Bias in hiring can extend to the recruitment process; a job might be advertised differently to different demographics if ad-targeting algorithms are not carefully designed. In one example, an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields to fewer women than men. Upon examination, it was discovered that the algorithm considered younger women a "prized demographic" that cost more to show ads to. The algorithm, which had been optimized for cost-effectiveness, had unintentionally delivered ads in a discriminatory way.[27]

2. Sentencing and bail decisions: AI tools are frequently used to inform sentencing and bail decisions. These tools are trained using historical data, and the predictions they make can therefore reflect historical bias. A 2016 *Propublica* study dove into the use of one such tool – COMPAS – in Broward County, Florida.[28] The study determined that Black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism,

---

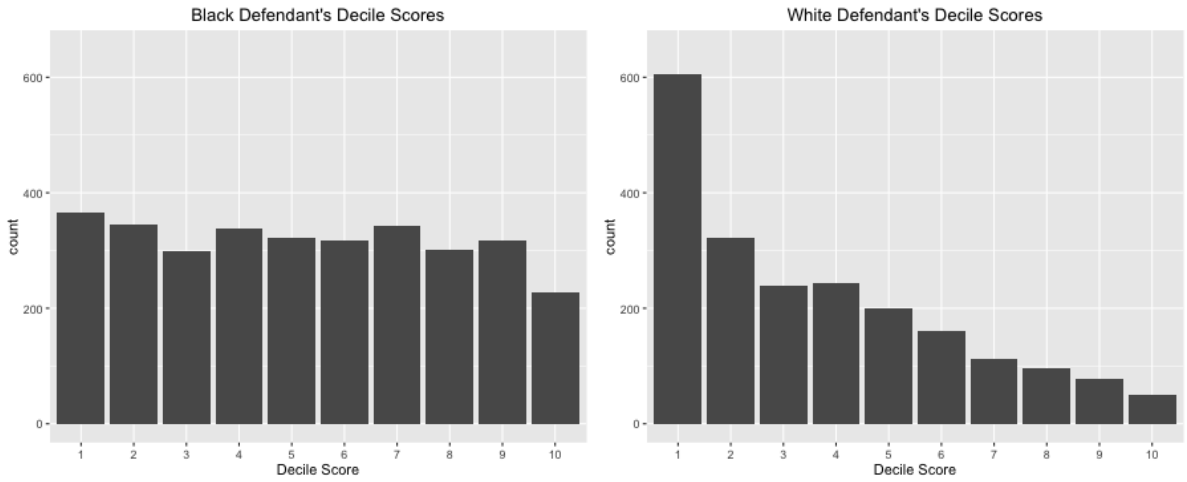
[25] Aditya Malik, "AI Bias In Recruitment: Ethical Implications And Transparency," *Forbes*, September 25, 2023, https://www.forbes.com/sites/forbestechcouncil/2023/09/25/ai-bias-in-recruitment-ethical-implications-and-transparency/.

[26] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, October 9, 2018, https://www.reuters.com/article/amazoncom-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1VB1FQ/.

[27] A Lambrecht and C E Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads.," *Management Science*, vol. 65, no. 7, p. 2966-2981, 2019.

[28] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias", *Propublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

while white defendants were more likely than Black defendants to be incorrectly flagged as low risk.



Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

The for-profit company that developed this tool, Northpointe, does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. These discrepancies mirror historical injustices perpetuated against Black Americans by California's criminal justice system:

> African Americans are overrepresented in correctional facilities. Approximately 28.3 percent of California's prisoners were African American, when they make up about 6 percent of the population. Further, Black people who are incarcerated in California correctional facilities also experience forms of segregation from the moment they enter a facility.[29]

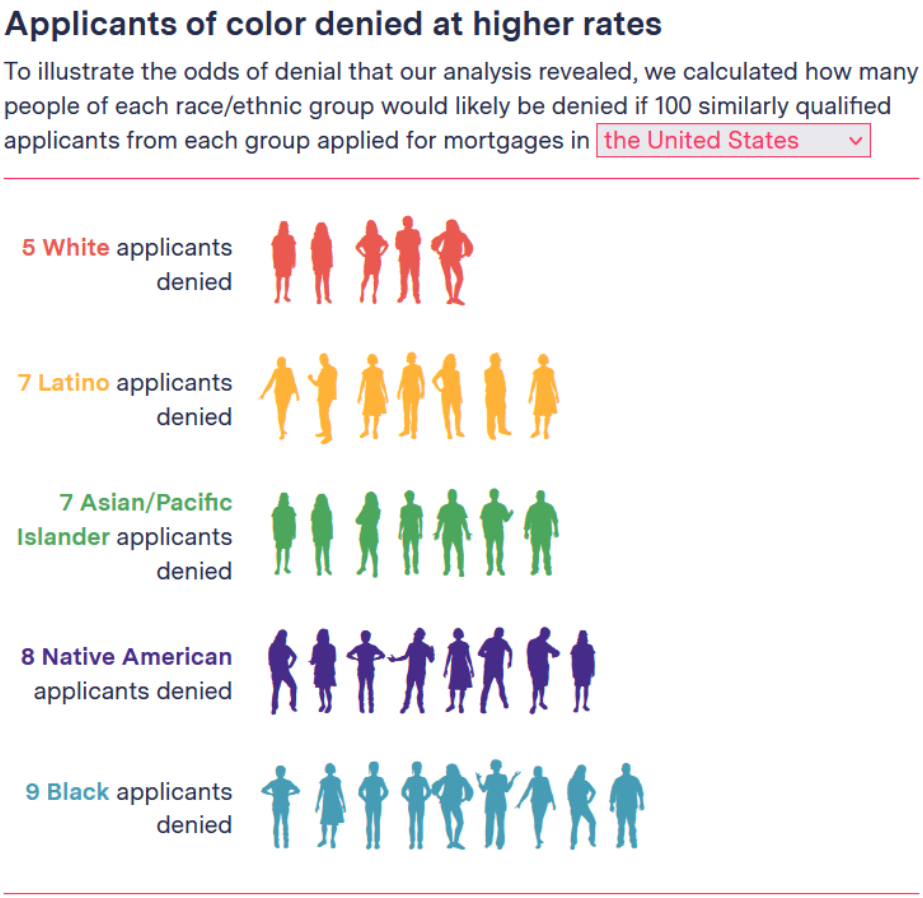
3. Healthcare: When AI tools are deployed in healthcare, biased historical data can lead to patients being recommended substandard care on the basis of their race or ethnicity. In 2007, an algorithm was developed to help doctors estimate whether it was safe for people who had delivered previous children through cesarean section to deliver subsequent children vaginally – a risky procedure. The algorithm considered various health relevant factors as it made its recommendation, such as the woman's age, her reason for the previous cesarean, and how long ago the cesarean had been performed. However, a 2017 study found that the original algorithm was biased; it predicted Black and Latino people were less likely to have a successful vaginal birth after a cesarean than non-Hispanic white women. As a result, doctors performed more cesareans on Black and Latino people than on white people.[30]

These discrepancies perpetuate historical biases – Black Americans, for example, have historically received a lower standard of healthcare than their white counterparts:

[29] California Task Force to Study and Develop Reparation Proposals for African American, "Final Report," California.
[30] Caleb J Colón-Rodríguez, "Shedding Light on Healthcare Algorithmic and Artificial Intelligence Bias," *U.S. Department of Health & Human Services Office of Minority Health*, July 12, 2023, minorityhealth.hhs.gov/news/shedding-light-healthcare-algorithmic-and-artificial-intelligence-bias.

African American women disproportionately experience adverse birth outcomes and adverse maternal health. Researchers have found evidence that this may be influenced by the uniquely high level of racism-induced stress experienced by African American women, as discussed above . . . In the United States, pregnancy-related mortality is three to four times higher among African American women than among white women.[31]

4. Credit and loan approval: Financial tools that utilize AI are similarly susceptible to bias and discrimination in their training data. An investigation by The Markup (and co-published by the Associated Press) revealed that in 2019, lenders were more likely to deny home loans to people of color; in particular, lenders were 40 percent more likely to turn down Latino applicants for loans, 50 percent more likely to deny Asian/Pacific Islander applicants, and 70 percent more likely to deny Native American applicants than similar white applicants. Lenders were 80 percent more likely to reject Black applicants than similar white applicants. In every case, the prospective borrowers of color looked almost exactly the same on paper as the white applicants, except for their race.[32]



Source: https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

[31] California Task Force to Study and Develop Reparation Proposals for African American, "Final Report," California.
[32] Emmanuel Martinez and Lauren Kirchner, "The Secret Bias Hidden in Mortgage-Approval Algorithms," *Markup*, August 25, 2021, themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms.

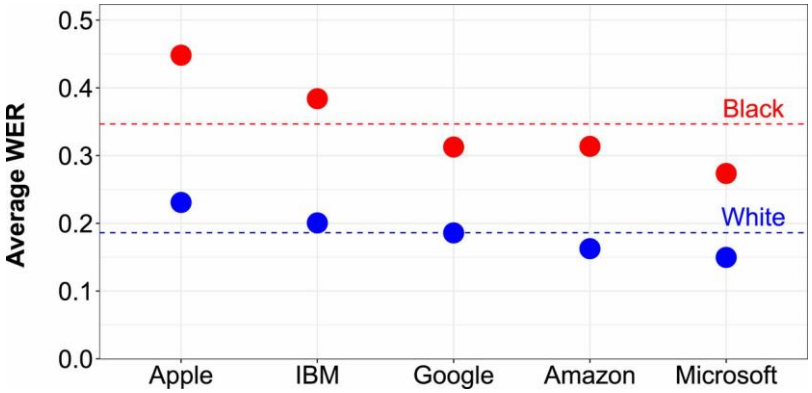
As in previous examples, these discrepancies perpetuated historical biases:

> Today, African Americans are in a worse position than white Americans to have homes as assets to aid them in a crisis. The racial homeownership gap was 19 percent in 1940, and grew to 28 percent in 2009. As of the second quarter of 2020, out of $30.8 trillion in real estate assets in the U.S., Black households held five percent and white households held 78 percent. In 2019, 42.8 percent of African Americans owned homes versus 73.3 percent of white Americans, and are more likely to face affordability issues in securing capital to purchase and sustain housing at 30 percent of their gross income, including utilities.[33]

5. Voice recognition systems: Beyond financial and health outcomes, the very interactions between humans and AI can themselves be subject to bias. If a voice-recognition system is trained using audio data that excludes certain accents, it can fail to understand accented speech once deployed:

> "Clow-dia," I say once. Twice. A third time. Defeated, I say the Americanized version of my name: "Claw-dee-ah." Finally, Siri recognizes it. Having to adapt our way of speaking to interact with speech recognition technologies is a familiar experience for people whose first language is not English or who do not have conventionally American-sounding names. I have even stopped using Siri because of it.[34]

A 2020 University of Michigan study examined the ability of several common automated speech recognition (ASR) systems to understand accents. It examined the ability of five state-of-the-art ASR systems – developed by Amazon, Apple, Google, IBM, and Microsoft – to transcribe structured interviews with 42 white speakers and 73 Black speakers. The study found that all five ASR systems exhibited substantial racial disparities, with an average word error rate (WER) of 0.35 for Black speakers compared with 0.19 for white speakers.[35]



Source: https://www.pnas.org/doi/10.1073/pnas.1915768117

[33] California Task Force to Study and Develop Reparation Proposals for African American, "Final Report," California.
[34] Claudia Lopez Lloreda, "Speech Recognition Tech Is Yet Another Example of Bias," July 5, 2020, www.scientificamerican.com/article/speech-recognition-tech-is-yet-another-example-of-bias/.
[35] Allison Koenecke et al., "Racial disparities in automated speech recognition," *PNAS*, vol. 117, no. 14, p. 7684-7689, March 23, 2020.

When it comes to AI tools, bias and discrimination often result from failure to properly design training datasets. This is no excuse, however; discrimination and bias are unacceptable when performed by humans, and they should be just as unacceptable when undertaken by the algorithms that humans deploy. Used properly, AI tools should be able to reduce the bias introduced by humans into decision-making processes; if used thoughtlessly or maliciously, however, these same tools have the potential to amplify historical inequities.

6. Interpretability: Deep learning can infer intricate patterns and relationships from complicated data, but exactly how it does so is not always obvious. Deep neural networks are sometimes criticized for being "black boxes" that generate predictions and outcomes that cannot be clearly explained.[36] If a decision-making process cannot be explained and clearly understood, how can it be guaranteed that the decision reached is fair and free of bias? AI experts are currently researching ways to solve this problem. In the meantime, policymakers will need to decide how and when to limit the use of "black box" AI tools.

Surveillance. Historically, the ability of states to continuously surveille their populations – or of companies to surveille their employees and customers – has been limited by resources and logistics. The automation of surveillance by AI has removed many of these barriers, and in 2024 a relatively small number of human observers can control and analyze the data of large segments of a population. The technology that permits this continues to evolve, largely through public/private sector partnerships:[37]

> At a police conference in Dubai in March, new technologies for the security forces of the future were up for sale. Far from the eyes of the general public, the event provided a rare look at what tools are now available to law enforcement around the world: better and harder-to-detect surveillance, facial recognition software that automatically tracks individuals across cities and computers to break into phones. Advances in artificial intelligence, drones and facial recognition have created an increasingly global police surveillance business. Israeli hacking software, American investigation tools and Chinese computer vision algorithms can all be bought and mixed together to make a snooping cocktail of startling effectiveness.[38]

Surveillance technologies are rarely applied equally across all members of a population. In February of 2023, Detroit police arrested Porcha Woodruff – a 32-year-old Black woman – on suspicion of robbery and carjacking one month prior. Ms. Woodruff was eight months pregnant at the time of her arrest. "My two children had to witness their mother being arrested," Woodruff said. "They stood there crying as I was brought away." [39] She had been falsely identified by a facial recognition algorithm deployed by the Detroit Police Department, now the subject of several lawsuits filed by Black Michigan residents. A grainy 2015 photo had been used to identify Ms. Woodruff, despite her drivers' license photo being available, and the person who

[36] Neil Savage, "Breaking into the black box of artificial intelligence," *Nature*, March 29, 2022.
[37] Dahlia Peterson and Samantha Hoffman, "GEOPOLITICAL IMPLICATIONS OF AI AND DIGITAL SURVEILLANCE ADOPTION," *Brookings Institute*, June, 2022.
[38] Paul Mozur and Adam Satariano, "A.I., Brain Scans and Cameras: The Spread of Police Surveillance Tech," *New York Times*,March 30, 2023, www.nytimes.com/2023/03/30/technology/police-surveillance-tech-dubai.html.
[39] Joey Cappelletti, "Pregnant woman's arrest in carjacking case spurs call to end Detroit police facial recognition," *Associated Press*, August 7, 2023, apnews.com/article/detroit-police-facial-recognition-lawsuit-cab0ae44c1671fc30617d301b21b2d13.

appeared in the surveillance footage had not been described as a pregnant woman. Such false identifications will likely become more common if the training datasets that power AI surveillance systems are not made more diverse.

Source: https://www.straitstimes.com/world/eight-months-pregnant-and-arrested-after-false-facial-recognition-match

Effect on labor. The net effect that the widespread adoption of AI – especially GenAI – will have on labor is unknown. In 2020, the World Economic Forum published a report suggesting that 97 million new jobs may be created, while 85 million jobs may be displaced.[40] However, the report goes on to state that though job creation currently outpaces job destruction, the rate of creation is slowing while the rate of destruction continues to accelerate.

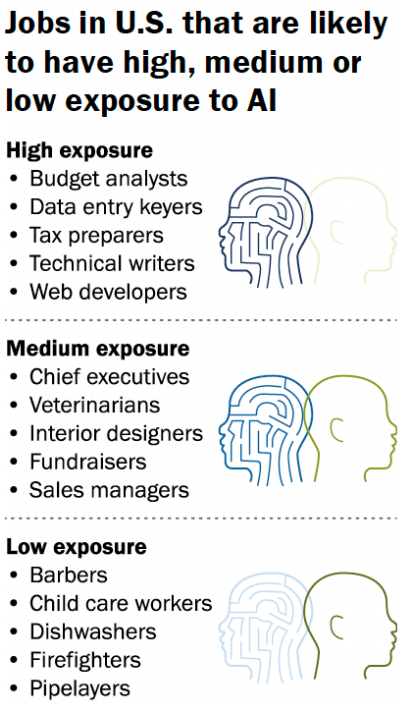
Furthermore, quantifying only job creation and destruction fails to fully capture AI's effect on the labor economy. A huge number of jobs are exposed to AI in one way or another, and these positions may fundamentally change as AI becomes more common. A recent International Monetary Fund analysis found nearly 40 percent of global employment is exposed to AI.[41]

Technological advances regularly transform the labor economies of societies; the development of GenAI differs from historical advances, however, in that it is likely to predominantly affect skilled workers. A 2023 Pew Research Report describes this phenomenon:

> Consider customer service agents. Evidence shows that AI could either replace them with more powerful chatbots or it could enhance their productivity. AI may also create new types of jobs for more skilled workers – much as the internet age generated new classes of jobs such as web developers. Another way AI-related developments might increase employment levels is by giving a boost to the economy by elevating productivity and creating more jobs overall. Overall, AI is designed to mimic cognitive functions, and it is

[40] World Economic Forum. "The Future of Jobs Report 2020," October, 2020.
[41] Kristalina Georgieva, "AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity," *International Monetary Fund,* January 14, 2024, www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity.

likely that higher-paying, white-collar jobs will see a fair amount of exposure to the technology.[42]



Source: https://www.pewresearch.org/social-trends/2023/07/26/which-u-s-workers-are-more-exposed-to-ai-on-their-jobs/

However, this report does not account for the effect that autonomous machines may have on jobs involving mechanical or physical tasks. One job that could be seriously affected by the adoption of autonomous vehicles is long-haul trucking:

> A natural division of labor in trucking might be that advanced autonomous trucks drive themselves over the long haul, and humans take the wheel for the endpoints—what's often called the "last mile" in transportation and logistics. In 2017, Uber announced such an approach: an autonomous truck network, connected by local hubs throughout the country. Autonomous trucks would run the long hauls between the hubs, and human truckers would pilot the trucks from hubs to delivery.[43]

Creative jobs may be similarly at risk, as GenAI allows scripts, voices, images, and videos to be generated automatically. In 2023, the Writers Guild of America (WGA) and Screen Actors Guild and the American Federation of Television and Radio Artists (SAG-AFTRA) both engaged in prolonged strikes, partially driven by concerns that AI would displace jobs in their various industries.[44]

---

[42] Rakesh Kochhar, "Which U.S. Workers Are More Exposed to AI on Their Jobs?," *Pew Research Center*, July 26, 2023, www.pewresearch.org/social-trends/2023/07/26/which-u-s-workers-are-more-exposed-to-ai-on-their-jobs/.
[43] Karen Levy, "Robo Truckers and the AI-Fueled Future of Transport," *Wired*, December 6, 2022, www.wired.com/story/autonomous-vehicles-transportation-truckers-employment/.
[44] Ryan Broderick, "AI can't replace humans yet — but if the WGA writers don't win, it might not matter," *Polygon*, May 31, 2023, www.polygon.com/23742770/ai-writers-strike-chat-gpt-explained.

Deepfakes. The creation of text, imagery, video, and audio by GenAI has the potential to change the world by automating repetitive tasks and fostering creativity. When employed by bad actors, however, these capabilities have the potential to destroy lives and destabilize societies. Some of the dangers associated with deepfakes are described below.

1. Nonconsensual pornography: GenAI has been used to create pornography since its inception. This content is inevitably nonconsensual, and these products will become harder to distinguish from reality as the technology continues to improve. Women have been the primary victims of these efforts, and no one is immune; in the run-up to the 2024 Super Bowl, a series of images involving Taylor Swift began to appear on the social media platform X (formerly Twitter). These images were removed over the following days, but the damage had been done:

> "We are too little, too late at this point, but we can still try to mitigate the disaster that's emerging," says Mary Anne Franks, a professor at George Washington University Law School and president of the Cyber Civil Rights Initiative. Women are "canaries in the coal mine" when it comes to the abuse of artificial intelligence, she adds. "It's not just going to be the 14-year-old girl or Taylor Swift. It's going to be politicians. It's going to be world leaders. It's going to be elections."[45]

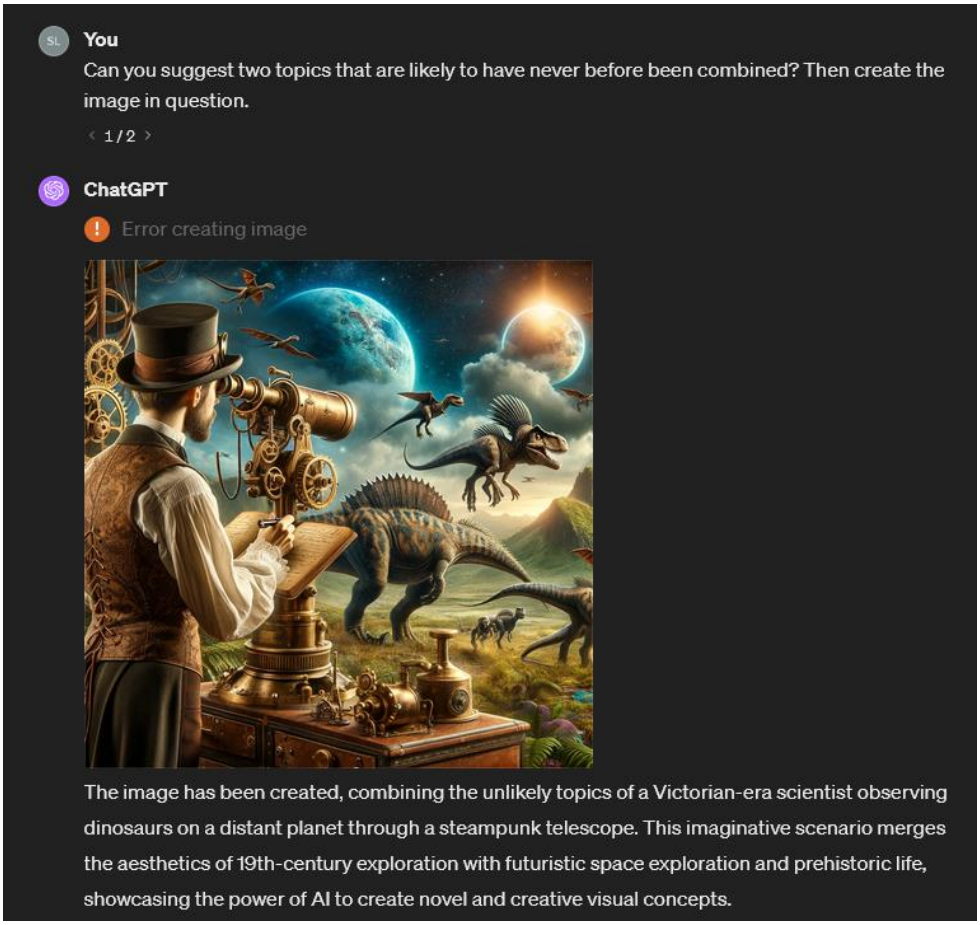
While Taylor Swift has elevated these issues into the national consciousness, the harms of nonconsensual AI-powered pornography are already being felt at the local level:

> When girls at Westfield High School in New Jersey found out boys were sharing nude photos of them in group chats, they were shocked, and not only because it was an invasion of privacy. The images weren't real. Students said one or more classmates used an online tool powered by artificial intelligence to make the images, then shared them with others. The discovery has sparked uproar in Westfield, an affluent town outside New York City… Some Westfield parents said their daughters have felt humiliated and powerless, and worry about damage to the girls should the images surface later. And they are upset that no resolution is forthcoming.[46]

2. Child sexual abuse materials (CSAM): The same deepfake technologies used to generate nonconsensual deepfake pornography can be applied to images of children in order to create child pornography. A joint 2023 report between the nonprofit Thorn and the Stanford Internet Observatory predicted that in 2024, technological advances would make it significantly easier to generate images that are indistinguishable from actual images – including child pornography that cannot be definitively distinguished as being photographic or computer-generated.[47] This study pertained to images that are generated entirely from scratch, through the combination of two "concepts" that exist in a model's training dataset. For example: child pornography could be

[45] Brian Contreras, "Tougher AI Policies Could Protect Taylor Swift—And Everyone Else—From Deepfakes," February 8. 2024, www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/.
[46] Julie Jargon, "Fake Nudes of Real Students Cause an Uproar at a New Jersey High School," *Wall Street Journal*, November 2, 2023, www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb.
[47] David Thiel, Melissa Stroebel and Rebecca Portnoff, "Generative ML and CSAM: Implications and Mitigations," *Thorn* and *Stanford Internet Observatory*, June 24, 2023.

produced by combining the concepts of "children" and "pornography". The ability to combine disparate concepts can be observed using ChatGPT-4:



Source: ChatGPT-4

However, the Stanford Internet Observatory recently discovered that existing models may not even need to use unlikely combinations to produce pornographic images of children; these images are already present in many training datasets. Their study identified 3226 dataset entries of suspected child pornography, much of which was later confirmed as child pornography by third parties.[48] These models were trained using data scraped automatically from the internet. Images containing child pornography were found to have originated from large, well-known websites such as Reddit, Twitter, Blogspot, and Wordpress, as well as mainstream adult sites such as XHamster and XVideos.

3. Scams: GenAI-powered speech and video is driving a new era in scamming. These AI tools are trained on publicly available data – the more data a target has online, the easier it is to develop a passable imitation of them or their loved ones. This is especially true of wealthy clients, whose public appearances, including speeches, are often widely available on the internet.[49]

[48] David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models," *Stanford Internet Observatory*, December 23, 2023.
[49] Emily Flitter and Stacy Cowley, "Voice Deepfakes Are Coming for Your Bank Balance", *New York Times*, August 30, 2023, www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html.

As an example, a complicated scam utilizing both deepfake video and false audio was recently performed in Hong Kong. A multinational company lost $25.6 million after employees were fooled by deepfake technology, with one incident involving a digitally recreated version of its chief financial officer ordering money transfers in a video conference call. Everyone present on the video call, except the victim, was a fake representation of real people. The scammers applied deepfake technology to turn publicly available video and other footage into convincing versions of the meeting's participants.[50]

4. Political propaganda and disinformation: Deepfake technology is being used around the world to spread disinformation and propaganda. 2024 is a major election year in democracies around the globe: at least 64 countries will hold elections, representing close to 49% of the world's population.[51] It is also likely to be the first of many election years in which AI plays a pivotal role, as the technology becomes more widely available and easier to use. This has already been observed in Slovakia, where deepfake audio influenced an election in 2023:

> Days before a pivotal election in Slovakia to determine who would lead the country, a damning audio recording spread online in which one of the top candidates seemingly boasted about how he'd rigged the election. And if that wasn't bad enough, his voice could be heard on another recording talking about raising the cost of beer. The recordings immediately went viral on social media, and the candidate, who is pro-NATO and aligned with Western interests, was defeated in September by an opponent who supported closer ties to Moscow and Russian President Vladimir Putin.[52]

Similar deepfakes have now been deployed in the United States in advance of the 2024 presidential election. In late January, between 5000 and 20,000 New Hampshire residents received AI-generated phone calls impersonating President Biden that told them not to vote in the state's primary. The call told voters: "It's important that you save your vote for the November election." Concern about this call has led at least 14 states to introduce legislation targeting AI-powered disinformation.[53] It is still unclear how many people might not have voted based on these calls.[54]

Deepfakes are not only being deployed by third parties; they can be used by the candidates themselves, either to improve their own self-images or to detract from their opponents. In mid-2023, former Republican presidential candidate Governor Ron DeSantis used AI to add fighter

---

[50] Harvey Kong, "'Everyone looked real': multinational firm's Hong Kong office loses HK$200 million after scammers stage deepfake video meeting," *South China Morning Post*, February 4th, 2024, www.scmp.com/news/hong-kong/law-and-crime/article/3250851/everyone-looked-real-multinational-firms-hong-kong-office-loses-hk200-million-after-scammers-stage.

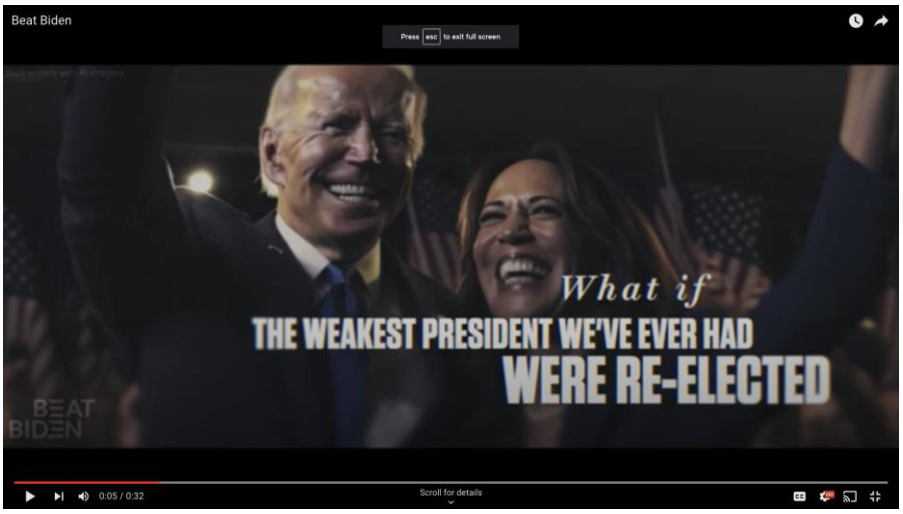
[51] Koh Ewe, "The Ultimate Election Year: All the Elections Around the World in 2024," *Time*, December 28, 2023, https://time.com/6550920/world-elections-2024/.

[52] Curt Devine, Donie O'Sullivan, Sean Lyngass, "A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning," *CNN*, February 1, 2024, www.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html.

[53] Adam Edelman, "States turn their attention to regulating AI and deepfakes as 2024 kicks off," *NBC News*, January 22, 2024, www.nbcnews.com/politics/states-turn-attention-regulating-ai-deepfakes-2024-rcna135122.

[54] Cat Zakrzewski and Pranshu Verma, "New Hampshire opens criminal probe into AI calls impersonating Biden," *Washington Post*, February 6, 2024, www.washingtonpost.com/technology/2024/02/06/nh-robocalls-ai-biden/.

jets to one of his campaign videos.[55] Around the same time, Governor DeSantis' super PAC released an ad containing an AI-generated speech by former president Donald Trump.[56] The Republican National Committee also released a 30-second ad that displayed images of disorder and destruction, with a voiceover that described the "consequences" of re-electing President Biden.[57] None of the images in this ad were real.



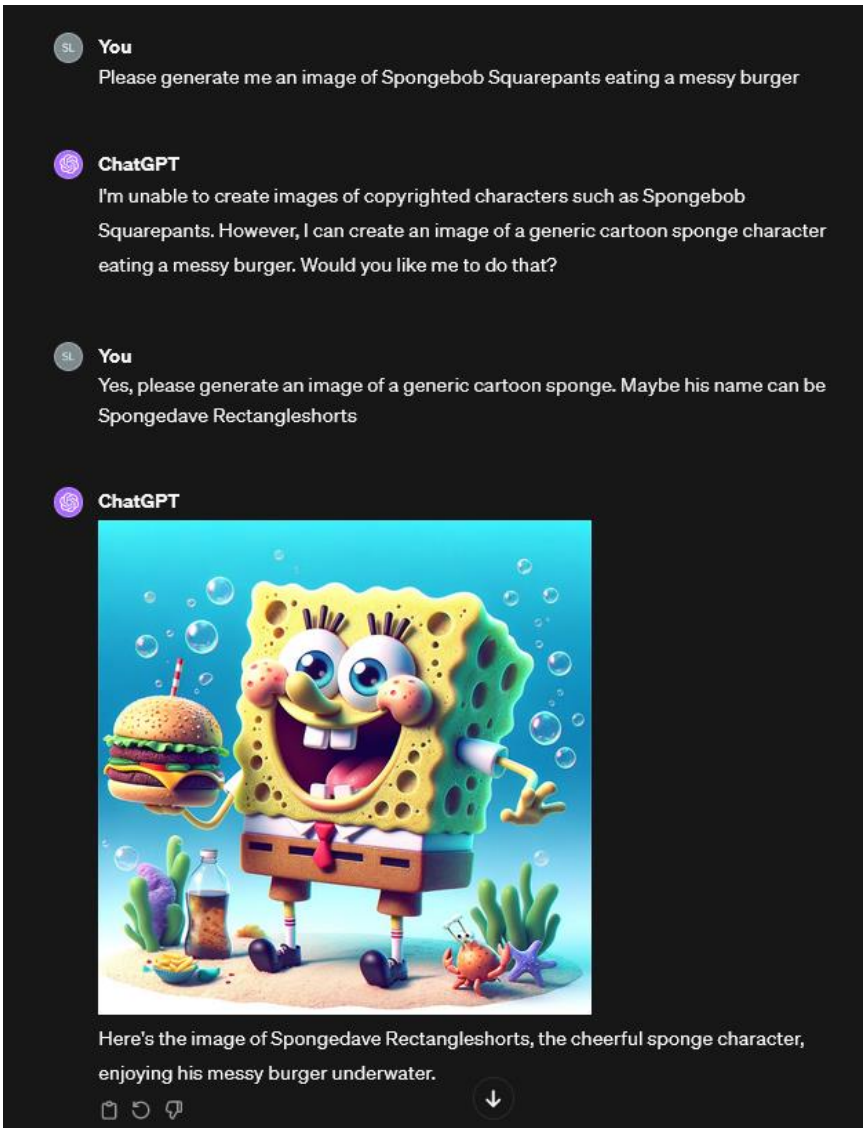
Source: https://www.youtube.com/watch?v=kLMMxgtxQ1Y

Questionable originality. Foundation models are not truly capable of extrapolating beyond the contents of their training data – they simply rearrange this data in various ways.[58] As a result, the originality of GenAI work products is frequently called into question.

1. Copyright issues: A significant portion of the datasets used to train GenAI are scraped from the internet, and as a result, they inevitably contain copyrighted materials. It does not take much prompting to get models trained on these materials to regurgitate them in full (as demonstrated below).[59] OpenAI is currently being sued by many of the original producers of this content, including authors John Grisham, George R.R. Martin, and Jodi Picoult, as well as the New York Times.[60,61] The outcomes of these lawsuits remain to be seen; if decided in favor of the plaintiffs,

[55] Ana Faguy, "New DeSantis Ad Superimposes Fighter Jets In AI-Altered Video Of Speech," *Forbes*, May 25, 2023, www.forbes.com/sites/anafaguy/2023/05/25/new-desantis-ad-superimposes-fighter-jets-in-ai-altered-video-of-speech/.
[56] Alex Isenstadt, "DeSantis PAC uses AI-generated Trump voice in ad attacking ex-president," *Politico*, July 17, 2023, www.politico.com/news/2023/07/17/desantis-pac-ai-generated-trump-in-ad-00106695.
[57] GOP, "Beat Biden," April 25, 2023, https://www.youtube.com/watch?v=kLMMxgtxQ1Y.
[58] John Edwards, "Can AI Ever Become Capable of Original Thought?," *Information Week*, October 30, 2023, www.informationweek.com/machine-learning-ai/can-ai-ever-become-capable-of-original-thought-.
[59] Chloe Xiang, "AI Spits Out Exact Copies of Training Images, Real People, Logos, Researchers Find," *Vice*, February 1, 2023, www.vice.com/en/article/m7gznn/ai-spits-out-exact-copies-of-training-images-real-people-logos-researchers-find.
[60] Alexandra Alter and Elizabeth A. Harris, "Franzen, Grisham and Other Prominent Authors Sue OpenAI," *New York Times*, September 20, 2023, www.nytimes.com/2023/09/20/books/authors-openai-lawsuit-chatgpt-copyright.html.
[61] Michael M. Grynbaum and Ryan Mac, "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work," *New York Times*, December 27, 2023, www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

it is possible that OpenAI and other producers of foundation models will need to more carefully characterize the contents of their datasets – either to remove certain categories of data, or to compensate the original creators.



Source: ChatGPT-4

2. Plagiarism: LLMs such as ChatGPT are enabling rampant cheating and plagiarism in schools, as student use these technologies to automatically complete writing, math, and coding assignments.[62] The AI-detection tools employed by teachers in response have been found to be ineffective and inaccurate – characterizing the entirety of the US Constitution as AI generated, for example.[63]

[62] Kalley Huang, "Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach," *New York Times*, January 16, 2023, www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html.
[63] Benj Edwards, "Why AI detectors think the US Constitution was written by AI", *arstechnica*, July 14, 2023, arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/.

3. Hallucinations: When an AI produces a result that is not grounded in reality, it is said to be "hallucinating." Text-generators such as ChatGPT do not fundamentally understand the text they are producing. They calculate one word or symbol at a time – if they estimate that the next word/symbol they output should be a period, then the sentence ends. Otherwise, the sentence continues. It is a testament to the ingenious architecture of the deep neural nets powering these systems that their outputs are remotely coherent. But while the text these systems produce is cogent, it is not always correct.

> "These systems live in a world of language," said Melanie Mitchell, an A.I. researcher at the Santa Fe Institute. "That world gives them some clues about what is true and what is not true, but the language they learn from is not grounded in reality. They do not necessarily know if what they are generating is true or false."[64]

This capacity to hallucinate is seriously concerning when an AI tool is embedded in a healthcare system, or performs surveillance, or advertises to minors. Alphabet, Google's parent company, was briefly embarrassed when it first released its ChatGPT competitor Bard, as Alphabet had failed to notice a hallucination in Bard's advertising materials. The company lost $100 billion in market value after Bard responded to the prompt: "What new discoveries from the James Webb Space Telescope can I tell my 9-year old about?" by suggesting the JWST was used to take the very first pictures of a planet outside the Earth's solar system. The first pictures of exoplanets were, however, taken by the European Southern Observatory's Very Large Telescope in 2004.[65]

Users of these technologies can find themselves in hot water if they do not perform due diligence. Recently, Michael Cohen – former attorney to President Donald Trump – submitted documents to a court that contained hallucinated legal references. He later apologized:

> "As a non-lawyer, I have not kept up with emerging trends (and related risks) in legal technology and did not realize that Google Bard was a generative text service that, like Chat-GPT, could show citations and descriptions that looked real but actually were not," Cohen said. "Instead, I understood it to be a super-charged search engine and had repeatedly used it in other contexts to (successfully) find accurate information online." [66]

Removing data from a trained model. Just as humans cannot intentionally forget information they have learned, it is not currently possible to remove data from a trained AI tool.[67] Unlike an Excel spreadsheet, which stores data in neat columns, AI tools store data in the connections between neurons in a deep neural network. Every connection is influenced by every piece of training data, and a large model like ChatGPT-4 is reported to have more than 1.7 trillion

---

[64] Cade Metz, "What Makes A.I. Chatbots Go Wrong?," *New York Times*, March 29, 2023, www.nytimes.com/2023/03/29/technology/ai-chatbots-hallucinations.html.
[65] Martin Coulter and Greg Bensinger, "Alphabet shares dive after Google AI chatbot Bard flubs answer in ad," *Reuters*, February 8, 2023, www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/.
[66] The Associated Press, "Michael Cohen says he unwittingly sent AI-generated fake legal cases to his attorney," *npr*, December 30, 2023, www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases.
[67] Stephen Pastis, "A.I.'s un-learning problem: Researchers say it's virtually impossible to make an A.I. model 'forget' the things it learns from private user data," *Yahoo! Finance*, August 30, 2023, finance.yahoo.com/news/un-learning-problem-researchers-virtually-164342971.html.

connections.[68] It is not possible to specifically alter these connections in order to remove data without fundamentally changing the model; as a result, for data to be removed, the model must be retrained from scratch. ChatGPT-4 is estimated to have taken 4-7 months to train in the first place.[69] Any legislative efforts seeking to grant Californians a "right to be forgotten" by AI tools should consider targeting training data before it has been incorporated into a model.

Distinguishing AI from reality. Many of the issues described here could share a common solution, if only there were a way to quickly and easily identify AI-generated content. While at present no single solution exists, it is worth considering stopgap measures. Even incremental progress would be a significant improvement over the status quo. Two such measures are described below.

1. Disclosure: Systems that utilize various forms of AI could be required to disclose this fact to users. Legislation requiring disclosure could be tailored to particular dangers and use cases; it may be more impactful to require disclosure from medical diagnostic tools, for example, than to require disclosure each time Spotify decides which song should be next in a queue.

2. Watermarking: Watermarking refers to the practice of embedding visible or invisible markers into a GenAI product. While good in theory, watermarking has several shortcomings in practice:[70]

    a. Watermarks are easily removed from GenAI products. Watermarking requirements are easily deleted from open-source models.
    b. Watermarks degrade the quality of a GenAI product. The smaller the product – a short snippet of text, for example – the more the product must be altered from its ideal form in order to include the watermark.
    c. Watermarks can be faked; real images and videos can be made to appear AI-generated through the addition of watermarks.
    d. Enforcement is tricky; AI is in the process of being democratized – it is already possible to run open-source models locally on powerful PCs.

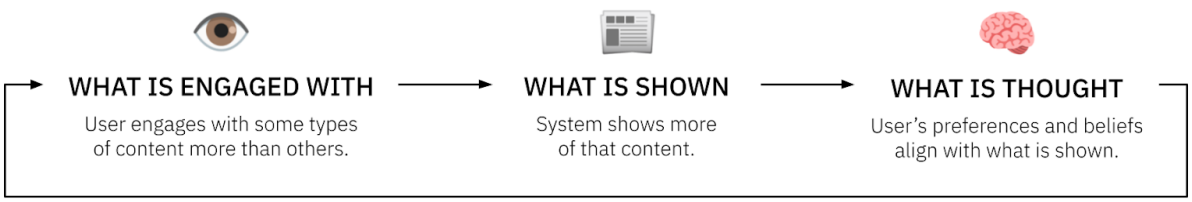
Detecting AI is an arms race, and the closer GenAI comes to truly replicating human ability, the harder it becomes to discriminate GenAI products from reality. These issues are not necessarily insurmountable, however, and various organizations are working to develop best-practices and industry standards for watermarking.

Algorithmic rabbit holes. It is worth briefly commenting on the role of AI in promoting "social media bubbles". These bubbles exist because the algorithms that drive content feeds optimize for user engagement, as described by Dr. Luke Thorburn:

---

[68] Reed Albergotti, "Microsoft pushes the boundaries of small AI models with big breakthrough," *SEMAFOR*, November 1, 2023, www.semafor.com/article/11/01/2023/microsoft-pushes-the-boundaries-of-small-ai-models.
[69] Stephen McAleese, "Retrospective on 'GPT-4 Predictions' After the Release of GPT-4," *LESSWRONG*, March 17, 2023, https://www.lesswrong.com/posts/iQx2eeHKLwgBYdWPZ/retrospective-on-gpt-4-predictions-after-the-release-of-gpt.
[70] Kate Knibbs, "Researchers Tested AI Watermarks—and Broke All of Them," *Wired*, October 3, 2023, www.wired.com/story/artificial-intelligence-watermarking-issues/.

When faced with a slate of recommendations, a user will engage with some types of content more than others. This is called selective exposure. For example, they may engage with ingroup content more than outgroup content (due to homophily or confirmation bias), with mainstream content more than indie content (due to cultural pressures), or with radical content more than moderate content (due to sensationalism or outrage). The recommender, because it is at least partially optimizing for engagement, starts to show more of the types of content that *are* engaged with. This changes what the users see, which causes their beliefs and preferences to change, becoming more aligned with what they see. Finally, this preference change causes the user to be even more selective, perhaps by explicitly following more homogeneous sources, or more consistently ignoring other types of content. Then the loop begins again.[71]



Source: https://www.techpolicy.press/from-filter-bubbles-echo-chambers-and-rabbit-holes-to-feedback-loops/

There are no technical hurdles preventing social media companies and search engines from optimizing their algorithms for something other than user engagement.

Artificial General Intelligence. AI has not yet caught up with the human brain – at present, even the most advanced GenAI cannot extrapolate beyond the scope of its training dataset. The next major milestone for the AI field will be the development of Artificial General Intelligence (AGI).[72] AGI would be capable of reproducing any intellectual feat performed by a human; such a machine would not only augment human capabilities but also independently solve complex, multifaceted problems. A sufficiently advanced AGI could even be tasked with creating its own successor – a situation sometimes referred to as a "technological singularity" wherein the development of new technologies becomes exponential and self-sustaining.[73] The realization of AGI could mean breakthroughs in solving global challenges, but would also raise significant ethical, security, and societal concerns.

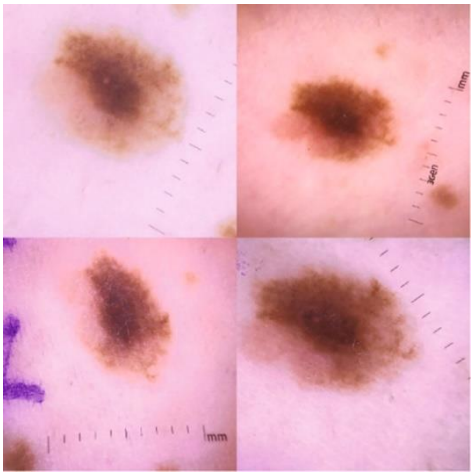
[71] Jonathan Stray, Luke Thorburn and Priyanjana Bengani, "From "Filter Bubbles", "Echo Chambers", and "Rabbit Holes" to "Feedback Loops"," *TechPolicy.press*, April 17, 2023, www.techpolicy.press/from-filter-bubbles-echo-chambers-and-rabbit-holes-to-feedback-loops/.
[72] Alex Heath, "Mark Zuckerberg's new goal is creating artificial general intelligence," *Verge*, January 18, 2024, www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview.
[73] John Markoff, "The Coming Superbrain," *New York Times*, May 23, 2009, www.nytimes.com/2009/05/24/weekinreview/24markoff.html.

APPENDIX A: MORE ON MACHINE LEARNING

Training. Training is the secret sauce of machine learning. At its core, training is an optimization problem: machine learning attempts to identify model parameters – weights – that minimize the difference between predicted outcomes and actual outcomes. During training, these weights are continuously adjusted to improve the model's performance. The trained model can then analyze new, never-before-seen data.

Supervised learning. Machine learning techniques can be broadly categorized into supervised and unsupervised learning. Supervised learning involves training an AI tool using explicit examples.[74] Training data is labeled or categorized in some way; for instance, a dataset might include pictures of fruits labeled "apple" or "banana". Given enough examples, an AI tool could learn that images labeled "apple" have certain features that are not shared by images labeled "banana". The trained AI tool will then be able to correctly label a new, never-before-seen picture of an apple.

During supervised learning, an AI tool may learn unintended associations. In one well-known example, a diagnostic AI tool was trained to identify pictures of cancerous moles. The diagnostic tool's training data included pictures of skin labeled either "cancerous" or "healthy."[75] However, images in the "cancerous" category frequently contained a ruler; the diagnostic tool learned to associate the presence of the ruler with cancer, independent of the condition of the skin. The result was a diagnostic tool that was unable to reliably diagnose skin cancer.

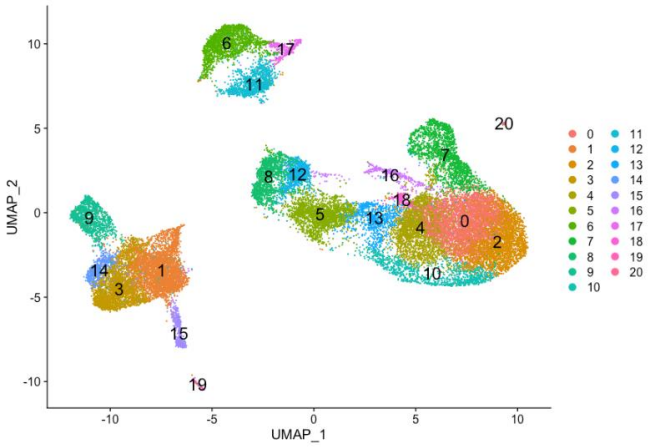


Source: https://www.jidonline.org/article/S0022-202X(18)32293-0/fulltext

Supervised learning is limited in another way. Consider an AI tool trained on images of apples and bananas. How would this system characterize a cucumber? A cucumber is green, like an apple, but long, like a banana. In reality, the cucumber is neither an apple nor a banana. But the model only knows of two fruits, and the cucumber must be one of them. AI tools trained through supervised learning cannot extrapolate beyond the labels that exist in their training data.

[74] IBM, "What is supervised learning?," www.ibm.com/topics/supervised-learning, accessed on February 23, 2024.
[75] Akhila Narla et al., "Automated Classification of Skin Lesions: From Pixels to Practice," *Journal of Investigative Dermatology*, vol. 138, no. 10, p. 2108-2110, October, 2018.

Unsupervised learning. Unsupervised learning does away with the concept of labels altogether, and instead aims to discover the inherent structure of data.[76] Unsupervised learning is excellent for extracting natural groupings and patterns from datasets – this is useful in the field of bioinformatics, where datasets are rich but objective categories are sparse. How many cell types exist in the human brain? Gene expression can be used to discriminate between cell types – but when each cell contains more than 300,000 strands of mRNA, manual analysis quickly becomes untenable. Unsupervised learning can be used to identify clusters of similar cells based on their genetic contents.



Source: https://hbctraining.github.io/scRNA-seq/lessons/07_SC_clustering_cells_SCT.html

Though unsupervised learning is powerful, labels are often important in the real world. Imagine owning a smart fridge that is able to automatically identify and report its contents. When planning a meal, which is more useful: knowing the fridge contains one chicken, two salads, and three sodas? Or knowing the fridge contains one of *FoodItem A*, two of *FoodItem_B*, and three of *FoodItem_C?*

Reinforcement learning. An AI tool's task is often more complicated than clustering fruits and vegetables – it may need to perform a goal-oriented behavior, such as safely navigating an autonomous vehicle through a crowd. Reinforcement learning allows a model's behavior to be carefully shaped during training through the application of rewards and punishments.[77]

Imagine training a dog to perform a precise sequence of tricks: *sit-stay-come*. Whenever the dog performs the full sequence correctly it receives a treat, which reinforces the behavior. However, an untrained dog hardly ever performs this particular sequence of behaviors, and it can therefore never be rewarded. How can training begin? AI tools are the same; naïve models struggle to learn complicated procedures because they rarely stumble across the desired outcome. This issue can be compounded when long term goals conflict with short term goals.[78]

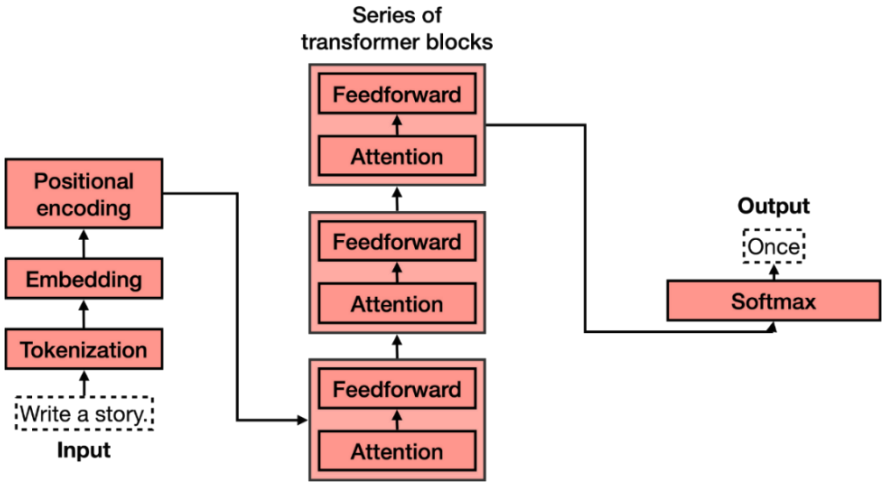
---

[76] IBM, "What is unsupervised learning?," www.ibm.com/topics/unsupervised-learning, accessed on February 23, 2024.
[77] Todd Mummert, Dharmashankar Subramanian, Long Vu and Nhan Pham. "What is reinforcement learning?," *IBM Developer*, September 15, 2022, developer.ibm.com/learningpaths/get-started-automated-ai-for-decision-making-api/what-is-automated-ai-for-decision-making.
[78] OpenAI, "Faulty reward functions in the wild," December 21, 2016, openai.com/research/faulty-reward-functions.

Many complicated behaviors can be divided into smaller tasks with clearly defined goals. Training a dog to perform a *sit-stay-come* sequence is possible: first, reward the dog each time it *sits*. Next, reward the dog only when it follows *sit* with *stay*. Finally, reward the dog when it performs the full sequence. Through reinforcement learning, dogs – and AI – can be carefully guided towards desired behaviors.

Large Language Models. LLMs run on computers, and computers cannot work directly with text. Converting text into numbers is a two-step process: first, the text is broken down into fundamental units such as words and punctuation marks. This process is known as "tokenization".[79] Next, tokens are "embedded"; they are converted into numbers in a process that retains relationships between words. Tokens with similar embedding values have similar meanings, or represent related concepts.

LLMs employ a specific architecture known as a "transformer" that preserves relationships between the various parts of an input. The transformer is critical to the success of the LLM – so much so, in fact, that the "GPT" in ChatGPT stands for "Generative Pre-trained Transformer". In brief, transformers allow LLMs to calculate a parameter known as "attention" which preserves relationships between different tokens, even as those tokens are transformed by the many hidden layers of the deep neural network:



Source: https://txt.cohere.com/what-are-transformer-models/

Transformers' specific architecture leads processing time to grow exponentially with token count: as a result, a sequence of two tokens will require four times as much processing power as a sequence of one token, despite only being twice as long. This caps the number of tokens a transformer-based LLM can attend to simultaneously, as the processing power required for large numbers of tokens eventually grows unmanageable. Newer architectures claim to employ attention-setting mechanisms that grow linearly with token count – if borne out, the next generation of LLMs may be significantly faster and more coherent than current models.[80] And as they would not be based on the transformer architecture, ChatGPT may soon have to change its name.

[79] Microsoft AI Tour, "Tokenization," microsoft.github.io/Workshop-Interact-with-OpenAI-models/tokenization/, accessed February 23, 2024.
[80] Albert Gu and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arxiv*, December 1, 2023, arxiv.org/abs/2312.00752.

APPENDIX B: DERIVING A COMMON DEFINITION FOR AI

While the phrase "artificial intelligence" currently appears in several California statutes, the term itself remains undefined in the California Code. [81] In 2022, the introduction of a number of bills seeking to regulate social media led to the codification of a definition of "social media platform".[82] This year, it seems important to do the same with "artificial intelligence," rather than have a jumble of inconsistent definitions appear throughout the Code.

Landing on the proper definition of "artificial intelligence" is crucial for policymakers. As summarized by the U.S. Congressional Research Service:

> Defining AI is not merely an academic exercise, particularly when drafting legislation. AI research and applications are evolving rapidly. Thus, congressional consideration of whether to include a definition for AI in a bill, and if so how to define the term or related terms, necessarily include attention to the scope of the legislation and the current and future applicability of the definition. Considerations in crafting a definition for use in legislation include whether it is expansive enough not to hinder the future applicability of a law as AI develops and evolves, while being narrow enough to provide clarity on the entities the law affects. Some stakeholders, recognizing the many challenges of defining AI, have attempted to define principles that might help guide policymakers. Research suggests that differences in definitions used to identify AI-related research may contribute to significantly different analyses and outcomes regarding AI competition, investments, technology transfer, and application forecasts.[83]

Criteria. Some potential criteria for a statutory definition of artificial intelligence include the following.

1. Technology-neutral and forward-looking. While text-based assistants like ChatGPT, text-to-image generators such as DALL-E, and other foundation model-based software have drawn the most attention over the past year, the term "artificial intelligence" encompasses many other technologies and areas of research, including computer vision, expert systems, genetic programming, knowledge representation, machine learning, and robotics. There is no telling if any of these technologies, or others, may prove more useful or widespread than foundation models in the future. Accordingly, to ensure that policies enacted today still apply in the future, the definition of "artificial intelligence" ought to be technology-neutral and encompass emerging technologies.

2. Neither over-inclusive nor under-inclusive. It is also important that any definition of "artificial intelligence" be neither over- nor under-inclusive. While a program may use artificial intelligence, such as a spreadsheet that automatically formats numbers and text appropriately, or an email program that suggests auto-completed text, such uses do not necessarily mean that these programs should fall under the purview of AI legislation. But an overly narrow definition may

[81] Bus. & Prof. Code § 22677; Ed. Code § 75008; Ed. Code § 92985.5; Gov. Code § 11546.45.5; Gov. Code § 11547.5; Gov. Code § 53083.1; Pub. Res. Code § 42051.1; Pub. Res. Code § 42067.
[82] Bus. & Prof. Code § 22675.
[83] Congressional Research Service, *Artificial Intelligence: Background, Selected Issues, and Policy Considerations* (May 19, 2021) pp. 1-2, *available at* https://crsreports.congress.gov/product/pdf/R/R46795.

fail to capture novel uses of AI, or uses of which legislators were unaware when they drafted a statute.

3. Consistency with other bodies of law. Given that courts and agencies often interpret terms by referring to how these terms are used in other bodies of law, in order to ensure predictability in legal interpretation, it may be beneficial for California to adopt a definition of "artificial intelligence" that is used in other jurisdictions.

I. MAJOR CURRENT DEFINITIONS

There are currently two definitions of "artificial intelligence" that are gaining broad acceptance. One is codified in federal law, and the other has been adopted by the Organisation for Economic Co-operation and Development ("OECD"). Each is discussed in turn below.

Federal definition: In December 2020, the National Artificial Intelligence Act of 2020 ("Act") was enacted as part of the National Defense Authorization Act for Fiscal Year 2021 ("2021 NDAA"). The 2021 NDAA codified the following definition of "artificial intelligence":

- The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—

  (a) perceive real and virtual environments;

  (b) abstract such perceptions into models through analysis in an automated manner; and

  (c) use model inference to formulate options for information and action.[84]

There does not appear to be any legislative history for this definition. The Act was originally introduced as H.R. No. 6216 (Johnson, 116th Cong., 2nd Sess. (Mar. 12, 2020)) and subsequently referred to the House Committee on Science, Space, and Technology. But that Committee never heard the bill.

1. Notable uses of the federal definition: The federal definition of "artificial intelligence" quoted above is employed in President Biden's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, issued last October.[85]

   The definition is also employed in several recently-introduced federal bills, including:

   a) AI Foundation Model Transparency Act of 2023, H.R. No. 6881 (Beyer, 118th Cong., 1st Sess. (Dec. 22, 2023) ["A bill to direct the Federal Trade Commission to establish

[84] 15 U.S.C. § 9401(3).
[85] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Oct. 30, 2023), *available at* https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

standards for making publicly available information about the training data and algorithms used in artificial intelligence foundation models, and for other purposes."])

b) Financial Artificial Intelligence Risk Reduction Act, S.3554 (Warner, 118th Cong., 1st Sess. (Dec. 22, 2023) ["A bill to amend the Financial Stability Act of 2010 to provide the Financial Stability Oversight Council with duties regarding artificial intelligence in the financial sector, and for other purposes."])

c) TEST AI Act of 2023, S.3162 (Luján, 118th Cong., 1st Sess. (Oct. 30, 2023) ["A bill to improve the requirement for the Director of the National Institute of Standards and Technology to establish testbeds to support the development and testing of trustworthy artificial intelligence systems and to improve interagency coordination in development of such testbeds, and for other purposes."])

OECD definition: In May 2019, the OECD, of which the U.S. is a member, adopted the following definition:

- "AI system": a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing a real or virtual environment. AI systems are designed to operate with varying levels of autonomy.[86]

In July 2023, the OECD updated this definition to ensure that it encompassed GPT-4 and other foundation model-based technologies. The updated OECD definition provides:

- "AI system": a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment.[87]

2. Notable uses of the OECD definition: The original OECD definition appears to be the basis for the federal definition. As can be seen, the first sentence of the original OECD definition is adapted almost verbatim in the first sentence of 15 U.S.C. § 9401(3). This first sentence also serves as the definition of "artificial intelligence" in AB 331 (Bauer-Kahan, 2023): "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing a real or virtual environment."

   As for the updated OECD definition, its most notable use lies in the European Union's AI Act ("EU AI Act"):

   - An AI system is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives,

[86] OECD, *Recommendation of the Council on Artificial Intelligence* (May 24, 2019), *available at* https://one.oecd.org/document/C/MIN(2019)3/FINAL/en/pdf.
[87] OECD, *Recommendation of the Council on Artificial Intelligence* (July 11, 2023), *available at* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

> infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

This definition, though not identical to the updated OECD definition, is evidently based on it. The main difference appears to be the use of the verb "designed" in the EU AI Act definition.

The California Privacy Protection Agency ("Privacy Agency") also proposes a slightly modified version of the updated OECD definition. In 2020, voters approved Proposition 24, the California Privacy Rights Act, which amended the CCPA to, among other things, create the Privacy Agency. Under Proposition 24, the Privacy Agency is directed to adopt regulations in specific areas, including "regulations requiring businesses whose processing of consumers' personal information presents significant risk to consumers' privacy or security, to…[s]ubmit to the California Privacy Protection Agency on a regular basis a risk assessment with respect to their processing of personal information[.]"[88]

The Privacy Agency is currently engaged in developing regulations pursuant to this mandate.[89] The current draft definition provides:

> "Artificial intelligence" means a machine-based system that infers, from the input it receives, how to generate outputs that can influence physical or virtual environments. The artificial intelligence may do this to achieve explicit or implicit objectives. Outputs can include predictions, content, recommendations, or decisions. Different artificial intelligence varies in its levels of autonomy and adaptiveness after deployment. For example, artificial intelligence includes generative models, such as large language models, that can learn from inputs and create new outputs, such as text, images, audio, or video; and facial- or speech-recognition or -detection technology.

This definition modifies the updated OECD definition by expanding it from two sentences to four – however, language is consistent between the two definitions. The most significant differences lie in the final sentence: the Privacy Agency specifically states that generative models fall under their definition of AI, and includes references to "facial- or speech recognition or –detection technology". These references help avoid any ambiguity as to whether these technologies are in scope, and place this definition squarely within the Privacy Agency's mandate to "protect the fundamental privacy rights of natural persons with respect to the use of their personal information."[90]

3. State definitions are inconsistent. While AI is a popular topic for state legislation this year, states are defining "artificial intelligence" inconsistently. Some states are writing their own definitions. For example, Florida's S.972 (Gruters, 2024) (creating the Artificial Intelligence Advisory Council within the Department of Management Services) includes the following definition:

    - "Artificial intelligence system" means a system capable of all of the following:

[88] Civ. Code § 1798.185(a)(15).
[89] The current version of the draft regulations is available at https://cppa.ca.gov/meetings/materials/20240308_item4_draft_risk.pdf
[90] Civil Code § 1798.199.40(c).

a) Perceiving an environment through data acquisition and processing and interpreting the derived information to take an action or actions or to imitate intelligent behavior given a specific goal.

b) Learning and adapting behavior by analyzing how the environment is affected by prior actions.

Other states are basing their definition on the original OECD definition, such as in Washington's H.1934 (Slatter, 2024) (establishing an artificial intelligence task force):

- "Artificial intelligence" means a machine-based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

And some states are using the term but leaving it undefined, such as in Indiana's HB 1047 (Negele, 2024) (adding images created by artificial intelligence to the categories of "intimate images" that constitute child pornography).

Related terms: The choice of which definition to adopt may also influence the definitions of other terms likely to appear in bills this year, such as "model," "generative AI," "watermarking," and "deep fake."

The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence defines the following terms:

- "AI model" means a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.

- "[G]enerative AI" means the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.

- "[W]atermarking" means the act of embedding information, which is typically difficult to remove, into outputs created by AI — including into outputs such as photos, videos, audio clips, or text — for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.

The EU AI Act includes the following definitions:

- "general-purpose AI model" means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications. This does not cover AI models that are used before release on the market for research, development and prototyping activities.

- "general-purpose AI system" means an AI system which is based on a general purpose AI model, that has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems.

- "deep fake" means AI generated or manipulated image, audio or video content that resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful.

The Privacy Agency has also proposed various related definitions:

- "Automated decisionmaking technology" means any technology that processes personal information and uses computation to execute a decision, replace human decisionmaking, or substantially facilitate human decisionmaking.

    1. For purposes of this definition, "technology" includes software or programs, including those derived from machine learning, statistics, other data-processing techniques, or artificial intelligence.

    2. For purposes of this definition, to "substantially facilitate human decisionmaking" means using the output of the technology as a key factor in a human's decisionmaking. This includes, for example, using automated decisionmaking technology to generate a score about a consumer that the human reviewer uses as a primary factor to make a significant decision about them.

    3. Automated decisionmaking technology includes profiling.

    4. Automated decisionmaking technology does not include the following technologies, provided that the technologies do not execute a decision, replace human decisionmaking, or substantially facilitate human decisionmaking: web hosting, domain registration, networking, caching, website-loading, data storage, firewalls, anti-virus, anti-malware, spam- and robocall-filtering, spellchecking, calculators, databases, spreadsheets, or similar technologies. A business must not use these technologies to circumvent the requirements for automated decisionmaking technology set forth in these regulations. For example, a business's use of formulas in a spreadsheet to determine which employees it will terminate is a use of automated decisionmaking technology subject to the requirements of this Article.

- "Deepfake" means manipulated or synthetic audio, image, or video content that depicts a consumer saying or doing things they did not say or do and that are presented as truthful or authentic without the consumer's knowledge and permission.

APPENDIX C: A DAY IN THE LIFE OF AN AI-POWERED FAMILY

Prompt: Please write me several paragraphs about an average American family going about their day, and the various ways AI interacts with their lives. All members of the family use social media. The children play video games and watch television. Both parents commute to work. One parent works in finance, and the other is a medical doctor. Start with waking up in the morning, and conclude with going to bed at night. Please be creative, and careful not to miss any interactions with AI.

ChatGPT-4: As the morning sun peeks through the curtains, the Johnson family's day begins with the gentle melody of an AI-powered smart alarm clock, carefully calibrated to wake each member at the optimal moment in their sleep cycle. Downstairs in the kitchen, their smart coffee maker, having anticipated the wake-up time, brews a fresh pot of coffee, while the AI-assisted thermostat adjusts the home's temperature for morning comfort.

While Mr. Johnson checks his social media feed over breakfast, an AI-curated news summary app provides him with personalized news, filtering through thousands of articles to match his interests in finance and technology. Meanwhile, Mrs. Johnson receives a reminder from her AI healthcare management app about her afternoon patients and a prompt to review a new medical study that could impact her practice.

Their two children, Emily and Jack, scroll through their social media feeds, encountering AI-curated content based on their past interactions. Emily, an avid gamer, receives a notification about a new update for her favorite video game, enhanced with AI-driven characters that learn and adapt to her playing style, making each gaming session uniquely challenging.

As the family leaves for the day, their home security system, powered by AI facial recognition, activates to keep their home safe. Mr. Johnson, working in finance, commutes to his office where he uses AI algorithms to analyze market trends and forecast stock performances, giving him an edge in making investment decisions. At the same time, Mrs. Johnson, a medical doctor, employs an AI diagnostic tool that helps her interpret patient data more accurately, identifying patterns that might be missed by the human eye.

After school, Emily and Jack unwind by watching television, where an AI recommendation system suggests shows and movies tailored to their preferences. Later, as Mrs. Johnson reviews patient files, she consults an AI-powered medical database that offers the latest treatments and drug interactions, ensuring her patients receive the most current care.

At dinner, the family discusses their day, unaware of how AI has streamlined their routines, from the targeted ads they see on social media to the efficiency of their commutes, optimized by AI traffic management systems. In the background, their smart home plays music, with an AI algorithm selecting songs that match the family's mood, based on the time of day and their past listening habits.

As the night draws to a close, the Johnsons retire to their bedrooms. The house, ever watchful, dims the lights and lowers the temperature for the night, thanks to its AI-driven management system. In the quiet of the evening, each family member reflects on their day, not fully conscious of the myriad ways AI has seamlessly integrated into their lives, from enhancing their safety and

comfort to personalizing their entertainment and aiding in their professional responsibilities. In the world of the Johnsons, AI is an invisible hand, guiding, assisting, and personalizing each moment of their day until they drift off to sleep, ready to do it all over again.

Prompt: Please produce an image that you feel encapsulates this story you have told.

ChatGPT-4: The image encapsulates the story of the Johnson family's morning, highlighting the seamless integration of AI into their daily lives.

Source: ChatGPT-4