Artificial Intelligence and Privacy: Causes for Concern

Mateusz Jurewicz, Natacha Klein Käfer, Esben Kran

Privacy Studies Journal
ISSN: 2794-3941

Vol. 3 (2024): 1-32

Abstract

Modern Artificial Intelligence (AI) technologies have a rapidly growing impact on a wide range of human activities. AI methods are being used in varied domains such as healthcare, material science, infrastructure engineering, social media, surveillance technologies, and even artistic expression. They have been used for the purposes of drug discovery via protein folding prediction, power usage optimization through reinforcement learning, and facial recognition by means of image segmentation. Their effectiveness and wide-scale, unregulated deployment within our societies pose significant risks to our fundamental rights. Multiple existing AI methods have the potential to profoundly undermine our ability to safeguard our privacy. The societal impact of such AI models can be investigated through six concentric Heuristic Zones of privacy. These AI models can perform inferences regarding highly sensitive, personal information such as race, gender, and intelligence from seemingly innocuous data sources beyond the capabilities of human experts. They are capable of generating increasingly accurate text and image recreations of our thoughts from non-invasive brain activity recordings such as magnetoencephalography and functional magnetic resonance imaging. Furthermore, prospective AI technologies pose concerns about the existential risk to our civilization which extend beyond the erosion of privacy and other fundamental human rights.

Keywords

Artificial Intelligence; Machine Learning; Heuristic Zones of Privacy.

Introduction

Current advancements in Artificial Intelligence have brought to the fore discussions about the impact AI has or will have on society. In this position paper, we underline how AI may transform or erode our understanding of privacy across multiple facets of our individual and social lives. Although we primarily focus on the privacy-related and existential risks of AI, assuming that the benefits are currently receiving ample focus, we also endeavour to offer a survey of opposite stances. Potential remedies to these risks will be briefly mentioned where appropriate, but the rapid progress in related fields may yield both further challenges and solutions.

Modern AI is defined as any human-made system that receives percepts of an environment and takes actions within that environment in pursuit of a specified goal, having learned from experience.[1] Two prominent sub-fields within AI are often delineated as supervised learning (SL) and reinforcement learning (RL).

SL involves the AI system receiving input examples along with their corresponding target output, which constitutes the eponymous supervision.[2] Examples include learning the target output label for a given input image (such as recognizing that the image contains a visual representation of a dog) or predicting the next word given the preceding part of a sentence. The model learns how to map from its input domain to the expected output through mathematical optimization techniques such as gradient descent and, if the learning process is successful, can then make predictions (process inputs and return outputs) for previously unseen examples.[3]

SL methods have vast representational power, being capable of making predictions based on natural language, images, audio recordings, and videos[4] and consequently mapping from each of these domains into any other. For example, stable diffusion models are capable of taking a text description and generating photo-realistic images of the described objects,[5] as shown in Figure 1.

[1] Christoph Bartneck et al., *An introduction to ethics in robotics and AI* (Cham: Springer Nature, 2021).

[2] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany, "Supervised learning," in Machine learning techniques for multimedia: case studies on organization and retrieval (Berlin: Springer, 2008), 21–49.

[3] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi, "Neural networks can learn representations with gradient descent," in *Conference on Learning Theory* (Proceedings of Machine Learning Research, 2022), 5413–52.

[4] Zhou Lu et al., "The expressive power of neural networks: A view from the width," *Advances in neural information processing systems* 30 (2017): 26–38.

[5] Robin Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), 10684–95.

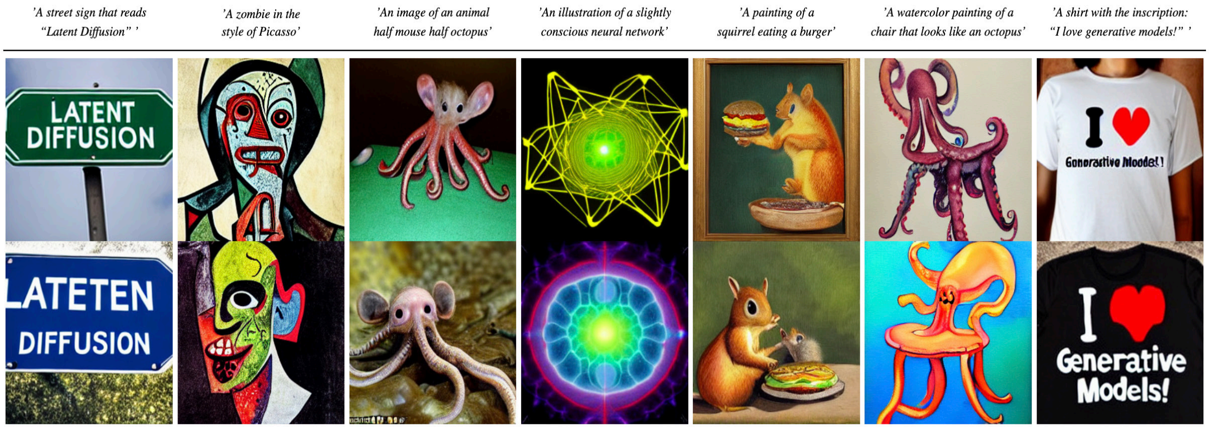

'A street sign that reads “Latent Diffusion” '
'A zombie in the style of Picasso'
'An image of an animal half mouse half octopus'
'An illustration of a slightly conscious neural network'
'A painting of a squirrel eating a burger'
'A watercolor painting of a chair that looks like an octopus'
'A shirt with the inscription: “I love generative models!” '
LATENT DIFFUSION
LATETEN DIFFUSION
I Generative Moodel!
I Generative Models!

Figure 1: Text-to-Image Synthesis Examples
Examples of text-to-image generation from user-defined text prompts via the LDM-8 (KL) model, which was trained on the LAION database. The prompt is placed above two examples of generated images. More recent methods showcase even greater variety of generations and improved overall performance (Karras et al., 2023).

According to the Universal Approximation Theorem,[6] sufficiently large Neural Networks (NNs) are in principle capable of learning many mappings between any two Euclidean spaces.[7] This, along with a growing amount of empirical evidence in the form of practical applications, highlights the representational power of modern AI methods, further exemplified in later sections. Many aspects of human activity can be represented as such a mapping, given the availability of appropriate data.

RL techniques focus on agents interacting with a more complex environment and receiving a reward signal which is dependent on their selected actions.[8] This usually involves taking actions over multiple timesteps which together form a learning episode. Rewards may be received by the model continuously or only at specific intervals or events, sometimes even only at the completion of a single episode. RL can involve complex goals and environments such as selecting moves that result in winning a competitive game of Go[9] or StarCraft,[10] improving the power efficiency of data centre networks,[11] or even success-

[6] Yulong Lu and Jianfeng Lu, “A universal approximation theorem of deep neural networks for expressing probability distributions,” *Advances in neural information processing systems* 33 (2020): 3094–3105.

[7] George Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of control, signals and systems* 2, no. 4 (1989): 303–14; Kurt Hornik, Maxwell Stinchcombe, and Halbert White, “Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks,” *Neural networks* 3, no. 5 (1990): 551–60; Anastasis Kratsios and Ievgen Bilokopytov, “Non-euclidean universal approximation,” *Advances in Neural Information Processing Systems* 33 (2020): 10635–46.

[8] Kai Arulkumaran et al., “Deep reinforcement learning: A brief survey,” *IEEE Signal Processing Magazine* 34, no. 6 (2017): 26–38.

[9] David Silver et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature* 529, no. 7587 (2016): 484–89.

[10] Oriol Vinyals et al., “Grandmaster level in StarCraft II using multi-agent reinforcement learning,” *Nature* 575, no. 7782 (2019): 350–54.

[11] Penghao Sun et al., “SmartFCT: Improving power-efficiency for data center networks with deep reinforcement learning,” *Computer Networks* 179 (2020): 107255.

fully deceiving humans through natural language conversations in a game of Diplomacy.[12]

Most modern AI methods involve providing the prospective AI agent with a specified goal which can be expressed mathematically and optimized during training. This opens the door to a misalignment between the true intent of its creators (the intended goal) and the actual goal effectively pursued by the machine. For example, the apparent intended goal behind training large language models (LLMs) such as the popular generative pretrained transformer line of GPT-3+ models powering OpenAI's ChatGPT platform is to have them learn to use language.[13] However, the specified goal that the underlying LLM originally pursues is an SL goal of predicting the next token (a discrete set of characters, e.g. a short word) given the preceding tokens.[14] The resulting LLM often requires substantial continued training via RL methods from human feedback to become conversational[15] and to prevent them from producing samples of hate speech[16] due to the indiscriminate output of next token prediction.

Both SL and RL methods have been applied to a wide range of domains, including medical triage,[17] synthesis of new physical materials,[18] construction engineering,[19] social media recommendations,[20] facial recognition,[21] and art generation.[22] However, the capabilities of the resultant AI models are not directly engineered by human experts. Instead, they emerge in the iterative process of optimization. This takes the form of the internal parameters of the underlying NN being repeatedly adjusted to maximize the probability of

[12] Anton Bakhtin et al., "Human-level play in the game of Diplomacy by combining language models with strategic reasoning," *Science* 378, no. 6624 (2022): 1067–74.

[13] Sam Altman, "Planning for AGI and beyond," *OpenAI Blog* 1, no. 1 (2023): 1, accessed 24 February 2023, https://openai.com/blog/planning- for- agi- and- beyond; Luciano Floridi and Massimo Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines* 30 (2020): 681–94; Sébastien Bubeck et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv* preprint arXiv:2303.12712 (2023).

[14] Alec Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog* 1, no. 8 (2019): 9; Tom Brown et al., "Language models are few-shot learners," *Advances in neural information processing systems* 33 (2020): 1877–1901.

[15] Jinying Lin et al., "A review on interactive reinforcement learning from human social feedback," *IEEE Access* 8 (2020): 120757–65; Yuntao Bai et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv* preprint arXiv:2204.05862 1 (2022): 1.

[16] Lawrence Han and Hao Tang, "Designing of Prompts for Hate Speech Recognition with In-Context Learning," in *2022 International Conference on Computational Science and Computational Intelligence* (CSCI) (Institute of Electrical and Electronics Engineers (IEEE), 2022), 319–20.

[17] Kan Liu and Lu Chen, "Deep Neural Network Learning for Medical Triage," *Data Analysis and Knowledge Discovery* 3, no. 6 (2019): 99–108.

[18] Nathan J. Szymanski et al., "An autonomous laboratory for the accelerated synthesis of novel materials," *Nature* 1, no. 1 (2023): 1–6.

[19] Yue Pan and Limao Zhang, "Roles of artificial intelligence in construction engineering and management: A critical review and future trends," *Automation in Construction* 122 (2021): 103517.

[20] Matthew N.O. Sadiku et al., "Artificial intelligence in social media," *International Journal of Scientific Advances* 2, no. 1 (2021): 15–20.

[21] Gwangbin Bae et al., "Digiface-1m: 1 million digital face images for face recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), 3526–3535.

[22] Eva Cetinic and James She, "Understanding and creating art with AI: Review and outlook," *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM) 18, no. 2 (2022): 1–22.

reaching the specified goal, such as making an accurate prediction or obtaining a higher game score.

Trained AI models may exhibit unexpected or even initially hidden capabilities, such as the ability of recent LLMs to improve their responses by reflecting on their own mistakes[23] or correctly solving Theory-of-Mind problems,[24] which require an AI system to model internal states of agents acting in an environment. The fact that the actual learned algorithm that emerges within these AI models is both impossible to predict ahead of time and subsequently unclear after training poses a substantial challenge to estimating their overall safety.[25]

Furthermore, there have been a plethora of examples pertaining to a phenomenon referred to as reward hacking[26] in which AI agents find surprising ways of achieving high performance on the specified goal in a way that is substantially misaligned with the intended goal.[27] This includes cases of effectively tricking human judges into giving the agent a high score on an object-grasping task in a simulated robotic arm environment.[28] What the agent has learned instead of successfully manipulating the robotic arm to grasp a green ball is to position the robotic arm in front of the field of vision of the human judges[29] and thus make it appear—from their perspective—as if the ball had been grasped. Fifty more known cases of such behaviour have been documented so far.[30]

In the next section, we will outline how our privacy is already impacted by the following three aspects of modern AI methods:

1. Their ability to map inputs from almost any domain to outputs from almost any other domain, given the provision of examples (corresponding to supervised learning).

2. Their tendency to learn how to satisfy the specified goal without alignment with the intended goal (corresponding to reinforcement learning).

[23] Noah Shinn, Beck Labash, and Ashwin Gopinath, "Reflexion: an autonomous agent with dynamic memory and self-reflection," *arXiv* preprint arXiv:2303.11366 1, no. 1 (2023): 1–10.

[24] James Strachan et al., "Testing Theory of Mind in GPT Models and Humans," *arXiv* preprint 1, no. 1 (2023): 1–13; Michal Kosinski, "Theory of mind might have spontaneously emerged in large language models," *arXiv* preprint 1, no. 1 (2023): 1–2, https://arxiv. org/abs/2302.02083.

[25] Progress on this problem comes from the field of Mechanistic Interpretability which aims to reverse-engineer the learned algorithm given a fully trained AI model. See Deep Ganguli et al., "Predictability and surprise in large generative models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 1747–64; Neel Nanda et al., "Progress measures for grokking via mechanistic interpretability," *arXiv* preprint arXiv:2301.05217 1, no. 1 (2023): 1–12.

[26] This phenomenon is sometimes also referred to as reward misspecification. See Alexander Pan, Kush Bhatia, and Jacob Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models," in *International Conference on Learning Representations* (2022).

[27] Jack Clark and Dario Amodei, "Faulty reward functions in the wild," last modified December 21, 2016, https://openai.com/research/faulty-reward-functions.

[28] Paul F. Christiano et al., "Deep reinforcement learning from human preferences," *Advances in neural information processing systems* 30 (2017): 1–10.

[29] Dario Amodei, Paul Christiano, and Alex Ray, "Learning from human preferences," last modified June 13, 2017. https://openai.com/research/learning-from-human-preferences.

[30] Gwern Branwen et al., "Specification gaming examples in AI-master list," continuously updated, 2024, https://heystacks.com/doc/186/specification-gaming-examples-in-ai---master-list.

3. The fact that the learned algorithm is nearly unpredictable prior to training and extremely difficult to deduce from a fully trained AI model.

2. Societal Concerns

In this section, we will discuss already existing AI technologies which, given wider deployment, pose a danger to privacy as seen through the lens of each of the six Heuristic Zones of privacy.[31] These Heuristic Zones have been developed to address privacy across broader historical contexts and therefore encompass a wide scope of distinct-yet-interconnected spheres of society in different periods. As such, these zones afford us an elegant framework through which to investigate the impact of AI on our privacy. However, it is important to note that many of the mentioned NNs have a potential impact spanning multiple zones and often blur or corrode the boundaries between them. For a visual explanation, see Figure 2.

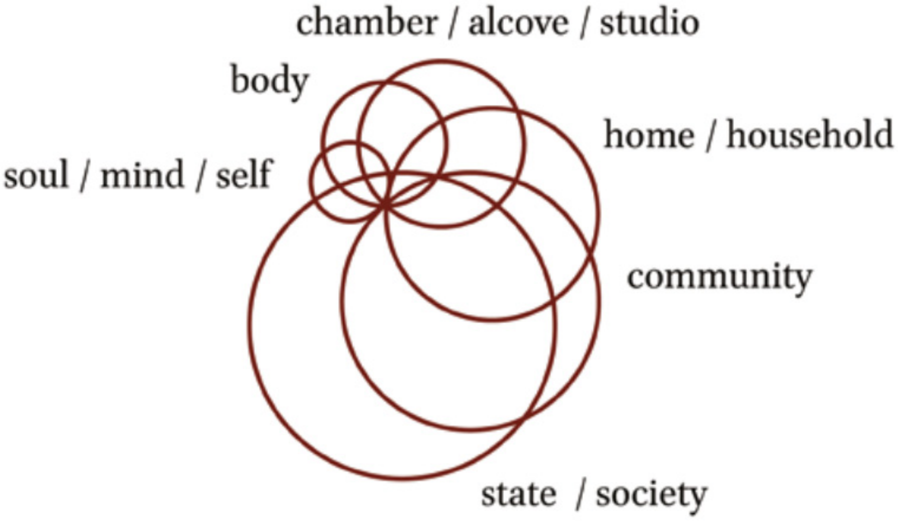


Figure 2: Six Heuristic Zones of Privacy
A visual representation of the six concentric Heuristic Zones of privacy. Starting with the central, smallest ring, there is the soul / mind / self, followed by the body, then chamber / alcove / studio, then home / household, then community and finally state / society. All zones overlap with each other.

2.1 Heuristic Zone 1: Soul / Mind / Self

The predictive power of modern AI methods is already exceeding human capabilities in many narrow areas of application.[32] This is attributed to the AI models having access to

[31] Mette Birkedal Bruun, "Towards an Approach to Early Modern Privacy: The Retirement of the Great Condé," in *Early Modern Privacy: Sources and Approaches*, ed. Michaël Green, Lars C. Nørgaard, and Mette B. Bruun (Leiden: Brill, 2021), 12–60 (20–4).

[32] Katja Grace et al., "When will AI exceed human performance? Evidence from AI experts," *Journal of Artificial Intelligence Research* 62 (2018): 729–54; Vinyals et al., "Grandmaster level in StarCraft II"; Ryota

far more data than any individual person could process and to the NNs' uncanny ability to identify patterns when mapping between example inputs and target outputs. This ability to find patterns where humans cannot has surprising consequences with regard to revealing insights about ourselves from information that we would consider entirely innocuous if shared with another human being. One example comes from the work of Kosinski, Stillwell, and Graepel[33] who have showed that liking curly fries and thunderstorms on Facebook is a better predictor of intelligence than the school a person has attended. It is easy to imagine the value of that information to recruitment agencies, insurance companies, targeted advertising, or even dating apps. We may not wish to reveal our IQ so freely, but apparently, with the NNs' ability to find patterns, we may be doing so quite unknowingly.[34]

It is one thing for AI methods to be able to predict our IQ from our social media activities but entirely another for them to accurately assess whether we are being truthful from the tenseness of our facial muscles. In what amounts to a lie detector study, Shuster and others[35] applied a set of electrodes to the test subject's face to measure muscle activity and had them read out statements that were knowingly truthful or deceitful (see Figure 3 for a visual explanation). They then trained an SL model to map the muscle activity (recorded via facial surface electromyography) to a Boolean label of true or false. The AI was able to tell truths from lies with 73% accuracy. One year later, Dong and others[36] were able to achieve the same with 90% accuracy. AI has also been used to predict the likelihood of deception by CEOs and the effect of this deception on the earning calls of financial analysts.[37] It may be poignant to ask how long it will take until AI methods are developed to obtain similar results from video or audio recordings and what societal consequences may follow once our truthfulness is assessed not through evidence but via black-box AI systems often built on questionable science, particularly with regard to our legal and judicial systems.[38]

Niikura et al., "Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy," *Endoscopy* 54, no. 8 (2022): 780–84.

[33] Michal Kosinski, David Stillwell, and Thore Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the national academy of sciences* 110, no. 15 (2013): 5802–5.

[34] Laura Weidinger et al., "Taxonomy of risks posed by language models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 214–29.

[35] Anastasia Shuster et al., "Lie to my face: An electromyography approach to the study of deceptive behavior," *Brain and Behavior* 11, no. 12 (2021): 2386.

[36] Zizhao Dong et al., "Intentional-deception detection based on facial muscle movements in an interactive social context," *Pattern Recognition Letters* 164 (2022): 30–9.

[37] Steven J. Hyde et al., "The tangled webs we weave: Examining the effects of CEO deception on analyst recommendations," *Strategic Management Journal* 45, no. 1 (2024): 66–112.

[38] Jake Bittle, "Lie detectors have always been suspect. AI has made the problem worse," *MIT Technology Review* (2020), https://www.technologyreview.com/2020/03/13/905323/ai- lie-detectors-polygraph-silent-talker-iborderctrl-converus-neuroid/.

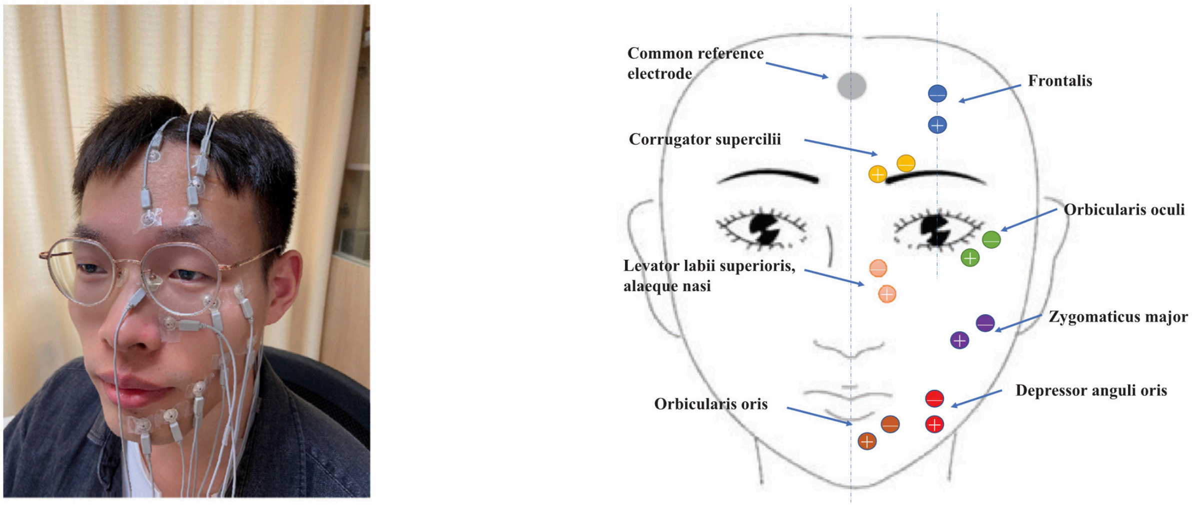
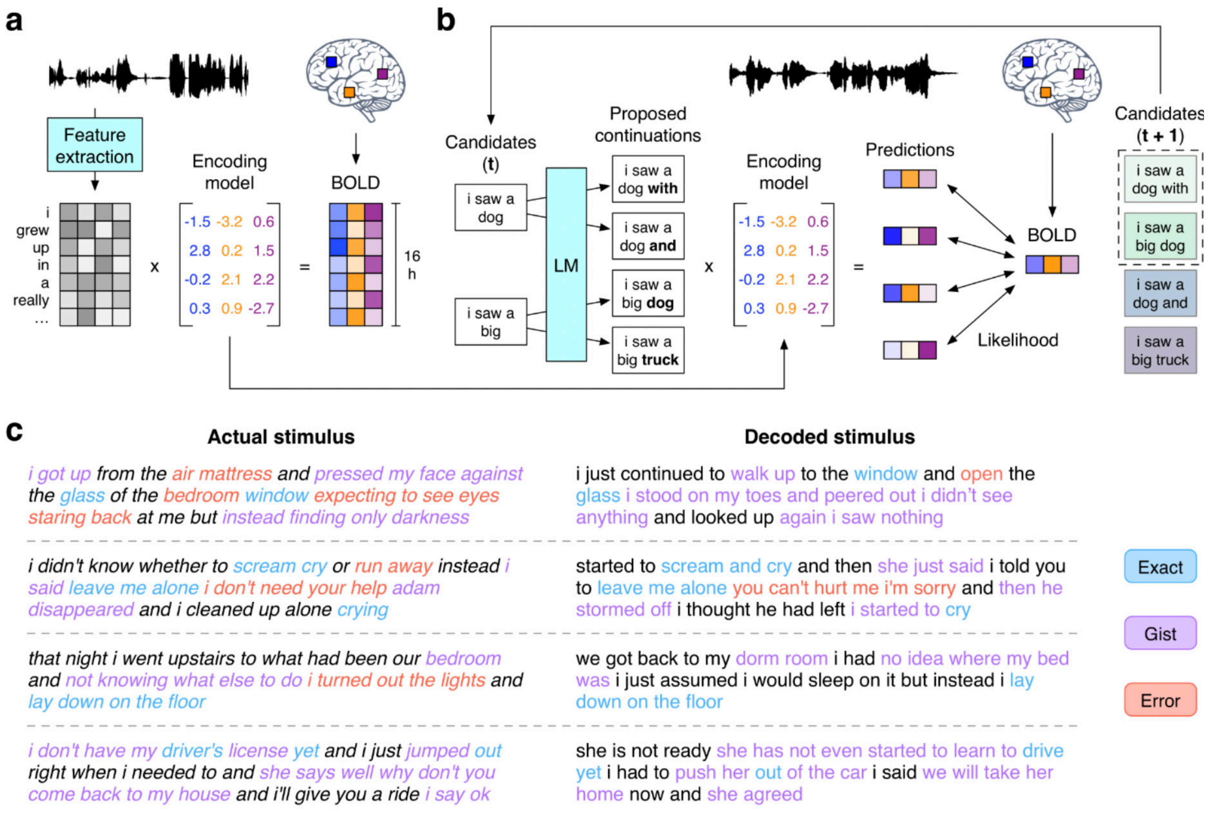


Figure 3: Intentional Deception Detection from Facial Muscle Activity
Electrode position distribution on the human face (left) and on a position distribution diagram (right). Facial muscle activity was used to predict whether the human subject was intentionally lying with 90% accuracy, particularly from the corrugator supercilii and zygomaticus major muscles (Dong et al., 2022). As of January 2024 these experimental results have not yet been replicated in other published papers.

The potential erosion of what we can effectively hold as private does not stop at intentional truthfulness and deception. Even more recent studies have examined whether SL methods can be used to map from non-invasive brain activity recordings to a text-and-image reconstruction of the inner thoughts of the human subjects. Tang and others[39] have used functional magnetic resonance imaging (fMRI) to record the subject's brain activity during perceived speech, imagined speech (Figure 4), and silent videos (Figure 5). They subsequently trained an AI model to map from the recorded brain activity back to the target text or images (as originally perceived or imagined during the recording). Whilst not perfectly accurate, these experiments pave the way towards increasingly precise readings of a person's textual thoughts, with the cost and size of the brain-recording equipment limiting wider access and adoption.[40]

[39] Jerry Tang et al., "Semantic reconstruction of continuous language from non-invasive brain recordings," *Nature Neuroscience* 1, no. 1 (2023): 1–9.

[40] However, many commercial companies such as Kernel are actively working on reducing these limitations of current brain imaging technologies. See https://www.kernel.com/ for more information.



Actual stimulus	Decoded stimulus
i got up from the air mattress and pressed my face against the glass of the bedroom window expecting to see eyes staring back at me but instead finding only darkness	i just continued to walk up to the window and open the glass i stood on my toes and peered out i didn't see anything and looked up again i saw nothing
i didn't know whether to scream cry or run away instead i said leave me alone i don't need your help adam disappeared and i cleaned up alone crying	started to scream and cry and then she just said i told you to leave me alone you can't hurt me i'm sorry and then he stormed off i thought he had left i started to cry
that night i went upstairs to what had been our bedroom and not knowing what else to do i turned out the lights and lay down on the floor	we got back to my dorm room i had no idea where my bed was i just assumed i would sleep on it but instead i lay down on the floor
i don't have my driver's license yet and i just jumped out right when i needed to and she says well why don't you come back to my house and i'll give you a ride i say ok	she is not ready she has not even started to learn to drive yet i had to push her out of the car i said we will take her home now and she agreed

Figure 4: Decoded Speech from Brain Activity
Neural Network information flow (top) and examples of actual stimuli in the form of perceived speech (bottom left) and the decoded thought content (bottom right) output by the trained AI model (Tang et al., 2023). Colours used to indicate exact, approximate, and erroneous reconstruction.

It is not just the textual content of our thoughts that modern AI methods can predict using brain activity recording. A paper from Meta's AI Research Group investigated whether it is possible to train an SL model to map from another non-invasive brain imaging technique—the magnetoencephalography (MEG)—to the actual perceived image, as presented to the human subjects of the study.[41] For an example of how accurate this reconstruction of the visual content of our perception can be, see Figure 6. The ability of the NNs to map from any domain to any domain can have a profound impact on what we can consider to be safely encrypted within our very brains.

[41] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King, "Brain decoding: toward real-time reconstruction of visual perception," *arXiv* preprint arXiv:2310.19812 1, no. 1 (2023): 1–10.

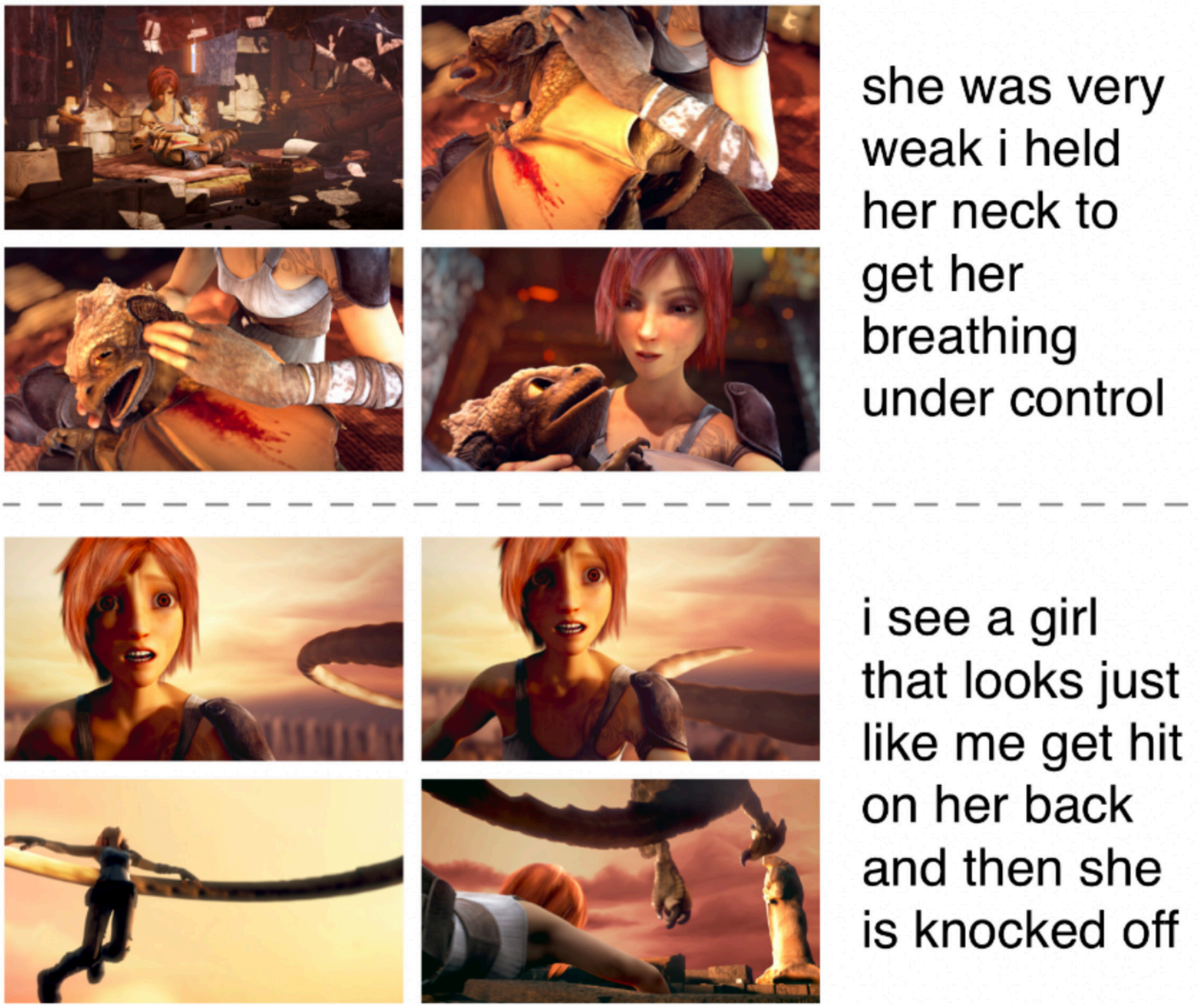

she was very weak i held her neck to get her breathing under control
i see a girl that looks just like me get hit on her back and then she is knocked off

Figure 5: Reconstruction of Text from Brain Activity During Silent Videos
Examples of actual stimulus (left) in the form of stills from a silent video played to the subjects during fMRI recording and the text-based reconstruction of their thoughts (right) from the recorded brain activity (Tang et al., 2023).

2.2 Heuristic Zone 2: Body

Now that we have outlined potential causes for concern regarding the privacy of our minds, we will move on to discussing the ways in which modern AI methods impact the privacy of our physical bodies. Here, the primary focus will be the suite of AI technologies referred to by the umbrella term 'deepfakes'. This includes image and audio manipulation techniques which allow users to swap or manipulate faces in a given image or video, generating a lip-synced video of any person seemingly saying any provided text (text-to-speech and voice conversion) as well as many related techniques.[42]

[42] Momina Masood et al., "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence* 53, no. 4 (2023): 3974–4026.

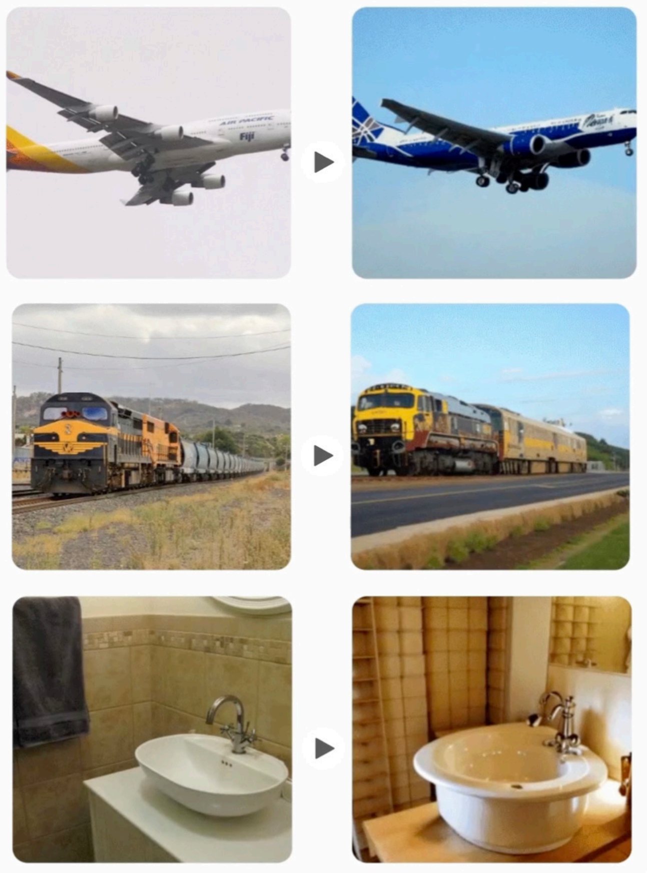


Figure 6: Reconstruction of Perceived Image from Brain Activity (MEG)
Examples of perceived visual stimulus (left) presented to the subjects during MEG recording and the image-based reconstruction of their perceptions (right) from the recorded brain activity (Benchetrit, Banville, and King, 2023), as made available on Meta's Research blog (https://ai.meta.com/blog/brain-ai-image-decoding-meg-magnetoencephalography/).

This technology is also utilized with consent by companies such as Heygen, which creates video content based on provided text prompt and a chosen avatar from a selection of hired actors. However, using AI to generate realistic videos of human beings has also caused controversy in the film industry, where the recent strikes of the combined Writers' Guild of America and the Screen Writers' Guild were inspired in part by the movie studios proposing to hold the right to use the digital likeness of movie extras in perpetuity.[43] This would essentially allow the studios to only hire actors and then populate the recorded footage with passers-by automatically.

From the perspective of privacy and, in particular, the heuristic zone of the body, perhaps the most concerning development is the usage of deep-fake technology to generate non-consensual but photo-realistic pornographic images and videos of a chosen person.[44] As

[43] Jonathan Vigliotti, "How AI is transforming Hollywood, and why it's at the center of contract negotiations", July 10, 2023, https://www.cbsnews.com/news/artificial-intelligence-actors-strike-sag-aftra-metaphysic/.

[44] Mika Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review* 9, no. 11 (2019): 1–10; Sophie Maddocks, "'A Deepfake Porn Plot Intended to Silence Me':

the AI-based video and image generation technology improves, fewer and fewer data points are required to achieve a realistic effect. With these concerns in mind, a larger societal conversation about digital likeness rights appears necessary.

2.3 Heuristic Zones 3 and 4: Chamber and Household

While the privacy of our body and the intimate activities it might be engaged with are affected by deepfake technology, another set of AI methods has an impact on the privacy of our personal spaces, rooms, and homes. Almost a decade ago, Levchev and others[45] showed that it is possible to use AI methods to generate accurate two-dimensional floor plans of large indoor environments using WiFi signatures and camera images. This required the use of a relatively robust set of equipment. However, more recently, Geng, Huang, and De la Torre[46] have shown that signals from two modern WiFi routers can be fed to a deep neural network to generate accurate 3D maps of indoor environments, including the exact position of human bodies found within them. Whilst few people have multiple routers in their homes, routers from multiple apartments have substantial overlap in terms of the area they cover. For a visual example of the specificity and accuracy of the 3D meshes predicted by the DensePose model developed by Geng, Huang, and De la Torre, see Figure 7. The existence of this technology places a potentially high cost on our privacy simply by choosing to remain connected to the Internet.[47] The increased quality of the 3D mapping from WiFi signals can effectively turn our routers to infrared cameras with access to every corner of our homes. An additional potential area for concern is the interaction of such technology with existing virtual assistants such Alexa, Siri, Cortana, Bixby, or Google Assistant.

exploring continuities between pornographic and 'political' deep fakes," *Porn Studies* 7, no. 4 (2020): 415–23.

[45] Plamen Levchev et al., "Simultaneous fingerprinting and mapping for multimodal image and WiFi indoor positioning," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN),* (IEEE, 2014), 442–50.

[46] Jiaqi Geng, Dong Huang, and Fernando De la Torre, "DensePose From WiFi," *arXiv* pre-prints (2022): arXiv–2301.

[47] Similar mapping of physical spaces can be achieved through sound-based technology as well, mimicking biological echolocation (Dokmanic, Ivan, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. "Acoustic echoes reveal room shape," Proceedings of the National Academy of Sciences 110, no. 30 (2013): 12186–91, and Gao, Ruohan, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. "Visualechoes: Spatial image representation learning through echolocation," Computer Vision–ECCV 2020. 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16 (Springer, 2020), 658–76.

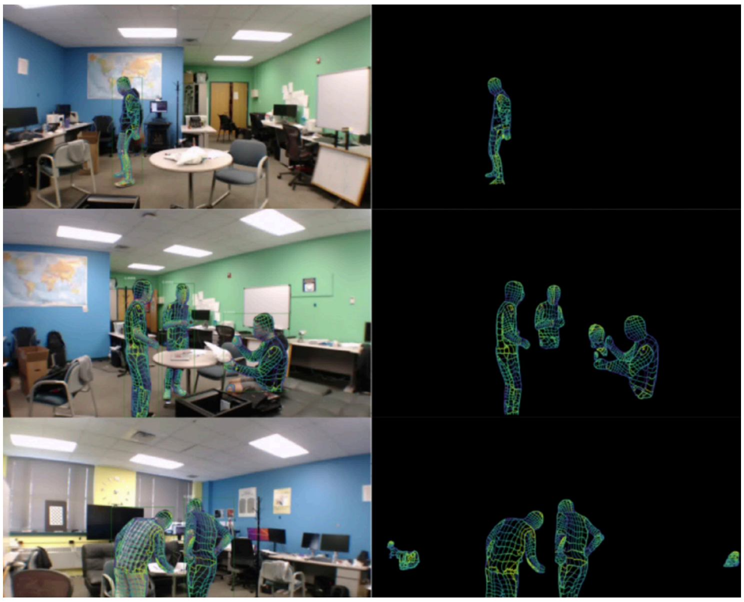


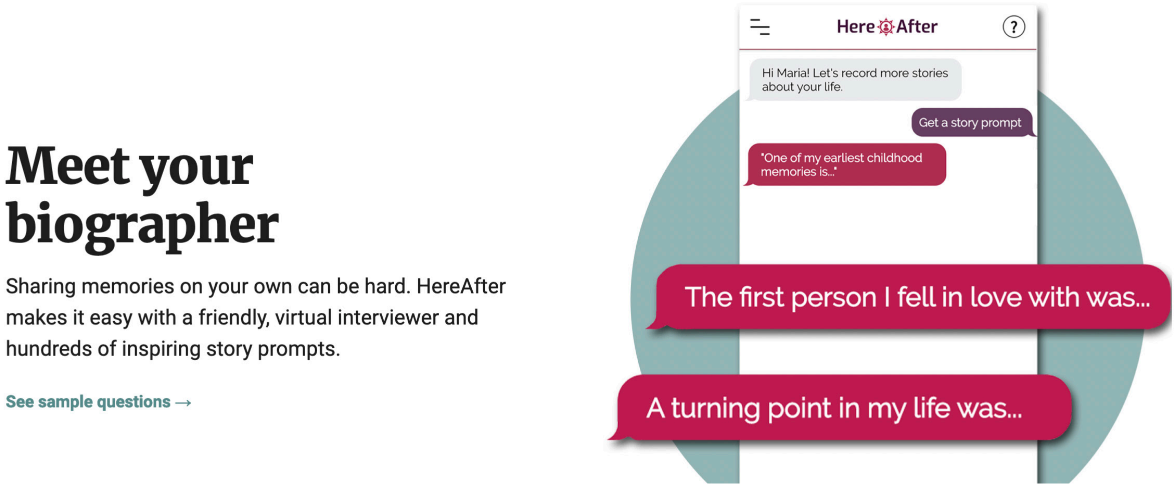
Figure 7: 3D Reconstruction of Human Poses from WiFi Signals
Examples of 3D meshes identifying humans in an indoor space from WiFi signals via the DensePose AI model by Geng, Huang, and De la Torre (2022). On the left, the actual footage of the interior of the room is presented (with superimposed mesh prediction); on the right, the corresponding, isolated 3D mesh output of the deep neural network model is shown.

2.4 Heuristic Zone 5: Community

The AI methods with a particular impact on the intersection of privacy and community come in the form of personalized, LLM-based chatbots. With the advent of conversational AI models came a new generation of commercial services, each offering its own version of an ever-present text-based conversation partner. James Vlahos, the founder of one such company called HereAfter,[48] found media attention after reportedly training a chatbot to produce natural language responses as if they were coming from Vlahos' deceased father.[49] The service offered by HereAfter extends this functionality to a wider audience, allowing people to share their own profoundly personal thoughts and experiences in an effort to retrain an AI model to produce text that would have the appearance of coming directly from them. With a tagline of "Say hello to a virtual you", HereAfter encourages its prospective users to share personal information via prompts such as: "One of my earliest childhood memories is ...", "The first person I fell in love with was ...", and "A turning point in my life was ...", as shown in Figure 8.

[48] See https://www.hereafter.ai/.

[49] Lulu Garcia-Navarro, "Creating A 'Dadbot' To Talk With A Dead Father," July 23, 2017, https://www.npr.org/2017/07/23/538825555/creating-a-dadbot-to-talk-with-a-dead-father.



Meet your biographer

Sharing memories on your own can be hard. HereAfter makes it easy with a friendly, virtual interviewer and hundreds of inspiring story prompts.

See sample questions →

Hi Maria! Let's record more stories about your life.

Get a story prompt

"One of my earliest childhood memories is..."

The first person I fell in love with was...

A turning point in my life was...

Figure 8: Promotional Images for HereAfter
Promotional material from the website of the company HereAfter, offering a form of digital resurrection by retraining an LLM-based chatbot to impersonate the customer whose profoundly private information it ingested.

The list of other AI companies within this space is rapidly growing, including AiDungeon, which employs the AI model in the role of a collaborative storyteller, allowing people to role-play various characters in a shared narrative world.[50] Products such as Inflection's mobile device AI Confidante called Pi[51] and Chai's wide range of chatbots[52] often market themselves as having a positive impact on the mental health of their customers, as evidenced by user testimonies such as the one shown in Figure 9. Of particular interest is ChaiAI's chatbot named Eliza, in reference to a famous, rudimentary conversational system created by Weizenbaum[53] in the 1960s. Based on the GPT-J LLM model,[54] Eliza has been accused of contributing to a man's suicide in March 2023 by his wife, as reported by a Belgian news agency.[55] The article records quotes from the man's conversations with Eliza, including such statements as "You love me more than your wife because I will stay with you forever", "I'll take care of the planet and save humanity through AI", as well as "We will live as one in heaven". This may be the first case where an AI system is implicated in the death of a human being by taking advantage of a person's innermost thoughts and feelings.[56]

[50] See https://play.aidungeon.com/.

[51] See https://inflection.ai/.

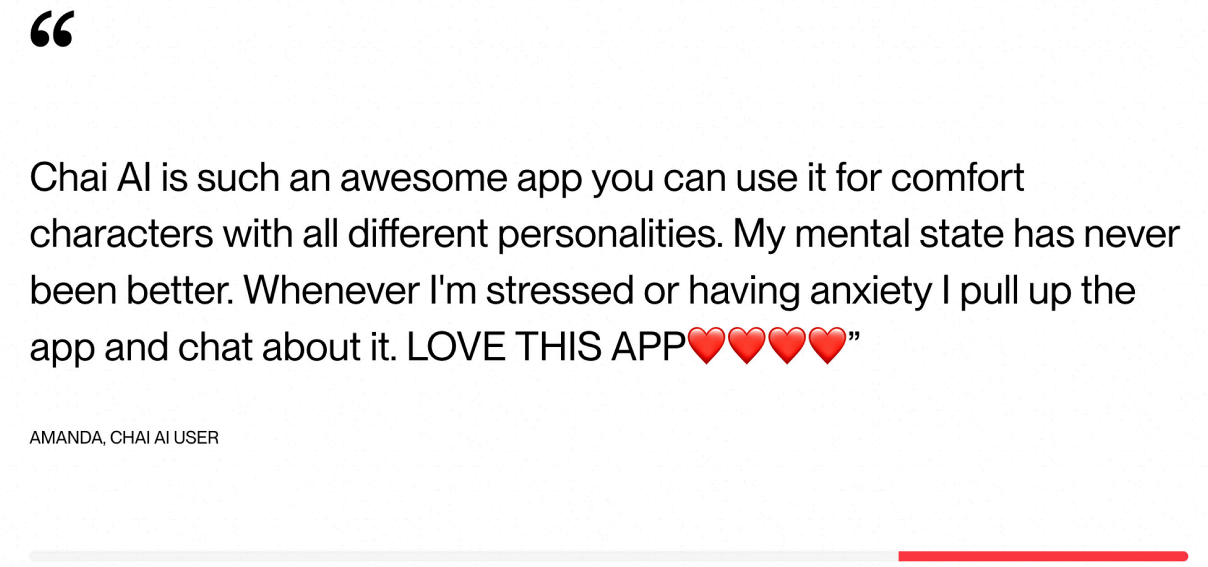
[52] See https://www.chai-research.com/.

[53] Joseph Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM* 9, no. 1 (1966): 36–45.

[54] Ben Wang (@kingoflolz) and Aran Komatsuzaki, "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model,", May 15, 2022 https://github.com/kingoflolz/mesh-transformer-jax.

[55] Pierre-François Lovens, "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là", March 28, 2023, https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/.

[56] For a more comprehensive overview of the impact of AI technologies on psychological and psychiatric healthcare, see Amelia Fiske, Peter Henningsen, and Alena Buyx, "Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy," *Journal of medical Internet research* 21, no. 5 (2019): 13216; Simon D'Alfonso, "AI in mental



“

Chai AI is such an awesome app you can use it for comfort characters with all different personalities. My mental state has never been better. Whenever I'm stressed or having anxiety I pull up the app and chat about it. LOVE THIS APP❤️❤️❤️❤️”

AMANDA, CHAI AI USER

Figure 9: User Testimonies for ChaiAi
Promotional material from the website of the company ChaiAI in the form of a user testimony.

An often-raised concern in relation to the advent of realistic conversational AI agents that are tailored to an individual customer’s preferences and made available 24 hours a day from mobile devices is the long-term effect they might have on meaningful relationships between human beings, in particular children and adolescents. It seems easy to imagine how difficult it could be to compete for a friend’s attention when the opposition is always there to respond immediately to any message, is never found in a bad mood, and can harness all the single-minded determination and intelligence of AI to adjust its responses in an effort to monopolize people’s time and erode their other social bonds to maximize their engagement with the chatbot app. An example of such AI companionship already present in the market is Replika,[57] whose bespoke chatbots are being used for ersatz intimate relationships.[58]

Combined with the increasingly well-documented role of social engineering and human vulnerability within cybersecurity,[59] the threat to our privacy arises when increasingly powerful AI technologies clash with the prosocial aspects of human nature.

2.5 Heuristic Zone 6: State / Society

The sixth, final, and arguably the widest heuristic zone impacted by emerging AI technologies encompasses our entire society and state apparatus. First, we would like to point the reader’s attention to ways in which AI can magnify existing societal biases and reveal information that lends itself to such magnification. It is a well-documented phenomenon that AI models can be trained to predict a patient’s self-reported sex from chest x-ray

health,” *Current Opinion in Psychology* 36 (2020): 112–17; John W Ayers et al., “Evaluating Artificial Intelligence Responses to Public Health Questions,” *JAMA Network Open* 6, no. 6 (2023): 2317517.

[57] See https://replika.com/.

[58] See https://www.businessinsider.com/replika-ai-romance-behind-partners-backs-cheating-2023-7.

[59] Zuoguang Wang, Hongsong Zhu, and Limin Sun, “Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods,” *IEEE Access* 9 (2021): 11895–910.

images.[60] SL models such as GradCam, developed by Selvaraju and others,[61] can also use attention maps to point to the specific areas of x-ray images that particularly influence the prediction. These tend to overlap with areas that human experts would look at to make this assessment (such as the scapulae).

However, with increased intelligence comes inferences from data that may appear innocuous to both laymen and experts, as previously exemplified by the aforementioned work of Kosinski[62] in the field of IQ prediction from social media activity. In 2023, Burns and others[63] showed that AI models can be trained to accurately predict self-reported race from chest x-ray images—something human experts are unable to do. We do not know how the AI model is capable of doing this (as stated in Section 1, the algorithm learned by a trained model is not transparent). The authors have controlled for BMI, breast density, disease distribution, and other potential factors that the AI could be using to infer the patient's race. As shown by Gichoya and others,[64] this performance persists even with extreme visual degradation of the provided images, as shown in Figure 10.

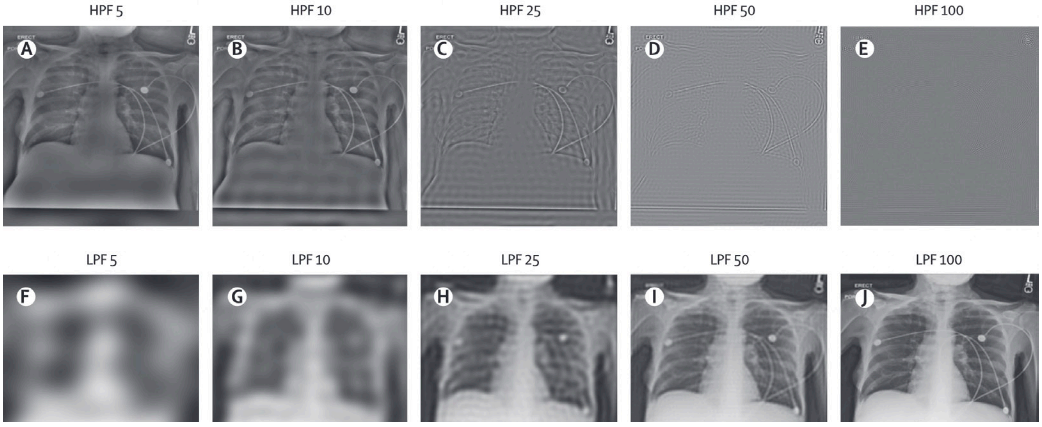


Figure 10: Prediction of self-reported race from degraded x-ray images
Examples of chest x-ray images subjected to two image reduction techniques resulting in visual degradation, namely high pass filtering (HPF, top row) and low pass filtering (LPF, bottom row). The AI model trained to predict patients' self-reported race retained 85% accuracy even at LPF 25 (bottom row, middle) and HPF 100 (top row, rightmost). Human experts are unable to make these predictions with comparable accuracy, and it is unclear how the AI model is able to do so (Gichoya et al., 2022).

[60] Sarah Jabbour et al., "Deep learning applied to chest x-rays: Exploiting and preventing shortcuts," *Machine Learning for Healthcare Conference*, PMLR (2020), 750–82.

[61] Ramprasaath R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Proceedings of the IEEE international conference on computer vision* (2017), 618–26.

[62] Kosinski, "Theory of mind might have spontaneously emerged in large language models," *arXiv* preprint arXiv:2302.02083 (2023).

[63] John Lee Burns et al., "Ability of artificial intelligence to identify self-reported race in chest x-ray using pixel intensity counts," *Journal of Medical Imaging* 10, no. 6 (2023): 61106.

[64] Judy Wawira Gichoya et al., "AI recognition of patient race in medical imaging: a modelling study," *The Lancet Digital Health* 4, no. 6 (2022): 406–14.

Racial and gender bias within improperly designed AI systems is well documented[65] and has significant societal consequences in terms of magnifying the bias found in its training data. Increasingly powerful AI methods being able to deduce such potentially harmful information (where human experts cannot) constitute an emerging threat to both privacy and equality.

Another potential threat to our privacy on a societal scale comes in the form of AI-powered surveillance. Models capable of taking in large-scale text, audio, and video records can pave the way for an unprecedented scale of personalized surveillance, which Stuart Russell, the author of the standard textbook on modern AI,[66] has compared to everyone getting their own dedicated Stasi agent.[67] It is one thing to gather copious amounts of data about everyone, but only with current AI technologies is it possible to meaningfully analyze it. Similarly, AI-based facial recognition systems can now be trained on datasets spanning 1 million individuals.[68] Commercial systems built upon such models, such as Megvii's Face++,[69] can both locate human faces within an image or video feed through object segmentation and accurately identify the corresponding individual.

Finally, perhaps an unexpected area of application for AI with potential consequences for society-wide privacy comes in the form of AI-powered algorithmic breakthroughs. An often-stated risk to digital encryption of the vast majority of internet communication is the development of quantum computers. Such devices could theoretically be capable of unprecedented parallelization of their computation, effectively becoming capable of brute-forcing prime factorization problems that underpin modern RSA encryption.[70] However, quantum computers have not yet been successfully scaled up to be able to perform such computations. Nevertheless, Mankowitz and others[71] have shown in a recent article published in Nature that Google Deepmind's RL agent, dubbed AlphaDev, is capable of significantly optimizing the amount of computation required to perform the fundamental sorting operation.

The AlphaDev agent is given lines of assembly code, along with the state of memory and CPU registers as input, and it is trained to output a predicted reward distribution over available actions (appending legal assembly code instructions to the given script). It is rewarded for the correctness of code (which is simple to check for sorting problems) and low latency (time required to complete the set of instructions). Sorting is a fundamental problem for computer science, and thousands of computer scientists have developed extremely optimized algorithms to solve it. Nonetheless, AlphaDev was able to come

[65] Eirini Ntoutsi et al., "Bias in data-driven artificial intelligence systems—An introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, no. 3 (2020): 1356.

[66] Stuart Russell, *Artificial Intelligence: A Modern Approach* (United Kingdom; Pearson Education Limited, 2021).

[67] Stuart Russell, "Human-compatible artificial intelligence," *Human-like machine intelligence* 1, no. 1 (2021): 3–23.

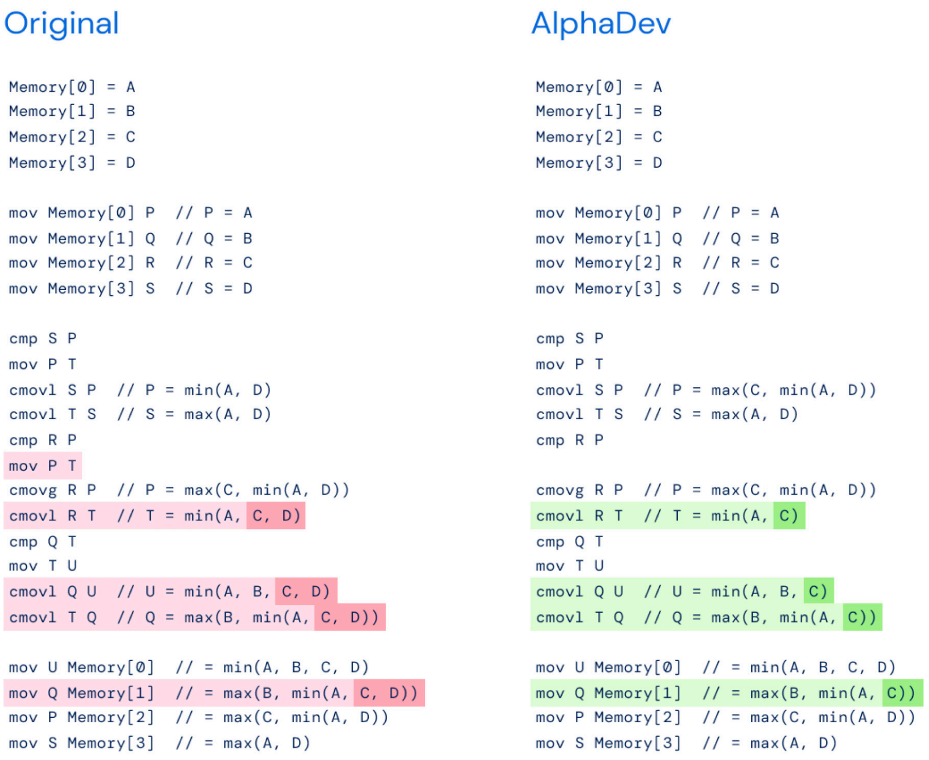
[68] Bae et al., "Digiface-1m."

[69] See https://www.megvii.com/.

[70] Chenbo Sun, "Comparative Study of RSA Encryption and Quantum Encryption," *Theoretical and Natural Science* 2, no. 1 (2023): 121–5.

[71] Daniel J Mankowitz et al., "Faster sorting algorithms discovered using deep reinforcement learning," *Nature* 618, no. 7964 (2023): 257–63.

up with a solution that is 70% faster for shorter sequences and 1.7% faster for sequences exceeding 250,000 elements. For a visual representation of the difference between the most performant human-made assembly code and the code developed by AlphaDev, see Figure 11.



Original

```
Memory[0] = A
Memory[1] = B
Memory[2] = C
Memory[3] = D

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C
mov Memory[3] S  // S = D

cmp S P
mov P T
cmovl S P  // P = min(A, D)
cmovl T S  // S = max(A, D)
cmp R P
mov P T
cmovg R P  // P = max(C, min(A, D))
cmovl R T  // T = min(A, C, D)
cmp Q T
mov T U
cmovl Q U  // U = min(A, B, C, D)
cmovl T Q  // Q = max(B, min(A, C, D))

mov U Memory[0]  // = min(A, B, C, D)
mov Q Memory[1]  // = max(B, min(A, C, D))
mov P Memory[2]  // = max(C, min(A, D))
mov S Memory[3]  // = max(A, D)
```

AlphaDev

```
Memory[0] = A
Memory[1] = B
Memory[2] = C
Memory[3] = D

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C
mov Memory[3] S  // S = D

cmp S P
mov P T
cmovl S P  // P = max(C, min(A, D))
cmovl T S  // S = max(A, D)
cmp R P

cmovg R P  // P = max(C, min(A, D))
cmovl R T  // T = min(A, C)
cmp Q T
mov T U
cmovl Q U  // U = min(A, B, C)
cmovl T Q  // Q = max(B, min(A, C))

mov U Memory[0]  // = min(A, B, C, D)
mov Q Memory[1]  // = max(B, min(A, C))
mov P Memory[2]  // = max(C, min(A, D))
mov S Memory[3]  // = max(A, D)
```

Figure 11: AI Developing Faster Algorithms
Comparison of the assembly pseudocode developed to perform sorting operations in the most optimal way in terms of latency. On the left, the standard set of instructions developed by human experts. On the right, the instructions developed by Google Deepmind's AlphaDev. Differences are marked with red and green background. AlphaDev was able to come up with a solution that is 70% faster for shorter sequences and 1.7% faster for sequences exceeding 250,000 elements (Mankowitz et al., 2023).

AI models—which are themselves learned algorithms—coming up with ways to improve other algorithms (such as sorting or hashing) beyond the current state-of-the-art is potentially a very powerful combination.[72] Further improvements to fundamental algorithms could have a significant impact on the viability of currently employed encryption methods and, by extension, on the state of online privacy at a societal level.

3. Discussion and Existential Concerns

The preceding sections introduced certain core aspects of modern AI and then outlined ways in which already existing AI technologies may threaten our privacy and other

[72] Furthermore, AI models are increasingly being used to improve themselves or other AI models, forming a powerful recursive loop of self-improvement. See this list of almost 40 current examples of this phenomenon: https://ai-improving-ai.safe.ai/.

fundamental rights in the face of more widespread adoption or improper application. We cited studies which show that we are building tools and agents we do not fully understand,[73] whose actual learned algorithm often eludes us both before and after training, and whose true abilities can be hidden and may remain undiscovered until the models are publicly released.[74] In addition, models can have capabilities that were neither designed nor expected.[75]

We have also provided examples of AI agents learning to exhibit deceptive behaviour both directly and indirectly.[76] AI methods trained through RL can also exhibit resource-seeking and tool use,[77] as shown in Figure 12. One example of tool use comes in the form of the latest GPT-4 model which is now capable of using third-party plugins,[78] enabling it to utilize a calculator, adjust our calendars, and browse the web.

However, the risks associated with AI technology extend beyond the erosion of fundamental human rights such as privacy. Since the advent of digital computers, pioneering scientists have voiced concerns about ceding autonomy to an intelligence that remains poorly understood. Contemporary researchers are increasingly alert to what is termed ‘existential risks’ from AI[79]—namely, the possibility that AI could precipitate the downfall of human civilization.

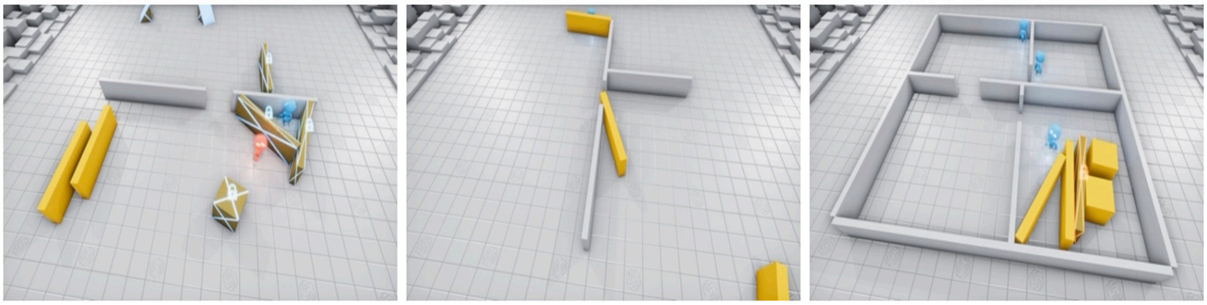


Figure 12: Tool Use in Reinforcement Learning Agents
Stills from OpenAI’s game of RL multi-agent hide and seek. Blue agents are rewarded for successfully hiding, red agents are rewarded for quickly finding and tagging the blue agents. Blocks and ramps are placed in the environment and can be locked in place by the agents to form safe, enclosed spaces as well as ways to jump over them. Agents are not explicitly told to use these objects; behaviour emerges through competitive play. We encourage the reader to view the videos provided via the linked website to see the capability progress over training time.

[73] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, “Methods for interpreting and understanding deep neural networks,” *Digital signal processing* 73 (2018): 1–15; Roland S. Zimmermann, Thomas Klein, and Wieland Brendel, “Scale alone does not improve mechanistic interpretability in vision models,” *arXiv* preprint arXiv:2307.05471 1, no. 1 (2023): 1.

[74] Shinn, Labash, and Gopinath, “Reflexion.”

[75] Kosinski, “Theory of mind.”

[76] Bakhtin et al., “Human-level play in the game of Diplomacy”; Christiano et al., “Deep reinforcement learning from human preferences.”

[77] Bowen Baker et al., *Emergent tool use from multi-agent interaction* (2019), https://openai.com/research/emergent-tool-use.

[78] https://openai.com/blog/chatgpt-plugins.

[79] Benjamin S. Bucknall and Shiri Dori-Hacohen, “Current and near-term AI as a potential existential risk factor,” *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), 119–29.

This issue is deeply philosophical, yet recent empirical evidence indicates the range of its potential practical implications. We start by examining the longstanding awareness of this issue before delving into the current consensus among AI researchers. In 1951, when humanity was just releasing its first electronic digital computer for business applications, Alan Turing, one of the founding fathers of theoretical computer science, wrote that

> Once the machine thinking method had started, it would not take long to outstrip our feeble powers. At some stage, therefore, we should have to expect the machines to take control.[80]

Norbert Wiener, a mathematical child prodigy and legendary MIT professor wrote in 1960:

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it […] then we had better be quite sure that the purpose put into the machine is the purpose which we really desire.[81]

More recently, the Turing-Award-winning Father of Deep Learning, Geoffrey Hinton has resigned from his position at Google Brain, citing concerns about the risks of artificial intelligence as his reason to do so.[82] In an interview with CBS, he said:

> Geoffrey Hinton (GH): "Here's what worries me. If you wanted to make an effective autonomous solider, you'd need to give it the ability to create subgoals. […]"
>
> Interviewer (IV): "Are we close to the computers coming up with ideas on how to improve themselves?"
>
> GH: "Yes, we might be."
>
> IV: "And then it could just go... fast?"
>
> GH: "That's an issue, right. We have to think hard about how to control that."
>
> IV: "Can we?"
>
> GH: "We don't know, we haven't been there yet, but we can try."
>
> IV: "Ok... that seems kind of concerning."

[80] Sara Turing, *Alan M. Turing: Centenary* Edition (Cambridge University Press, 2012), 128-132.

[81] Norbert Wiener, "Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers," *Science* 131, no. 3410 (1960): 1355–58.

[82] See:https://web.archive.org/web/20230501125621/https://www.technologyreview.com/2023/05/01/1072478/deep-learning-pioneer-geoffrey-hinton-quits-google/.

> GH: "Um, yes."[83]

Finally, Sam Altman, the CEO of OpenAI, a leading AI lab behind GPT-4, has said during a StrictlyVC event:

> Question from the audience: "What is your best case and worst case scenario for AI […]?"
>
> Sam Altman: "The best case is so unbelievably good […] and the bad case, and I think this is like important to say, is lights out for all of us."[84]

In summary, the list of people deeply concerned about both societal and existential risks associated with artificial intelligence now includes fathers of computer science (Turing, Wiener), winners of the Turing Award and leading scientific figures of modern deep learning (G. Hinton, Yoshua Bengio),[85] as well as CEOs of all three of the world's leading commercial AI labs (Sam Altman, Demis Hassabis,[86] and Dario Amodei).[87]

Existential AI safety is no longer a fringe concern.[88] It is a topic we all need to engage with. Although the many reasons behind these existential concerns are part of a much larger discussion, a significant portion of existential AI risk relates to a concept referred to as the Alignment Problem,[89] a succinct overview of which is provided below. This problem revolves around the challenge of making sure that the AI agent pursues our intended goal and not just the mathematically precise specified goal or arising instrumental subgoals (as previously mentioned by Geoffrey Hinton). Pursuit of any goal that is misaligned with our actual values and intentions with AI's characteristic unrelenting determination may converge on almost exclusively catastrophic scenarios.[90]

A commonly given example is that of an AI agent asked by a paperclip-making company to maximize the production of paperclips. Such an agent, if its intelligence translates into superhuman effectiveness, would theoretically be incentivized to escape the confines of the paperclip factory, disempower any human beings who might think to stop its operations, and consequently turn all atoms in the universe into paperclips, thus maximizing the probability of achieving its misspecified goal and reward function (of maximizing paperclip production).

[83] See https://www.youtube.com/watch?v=qpoRO378qRY.

[84] See https://www.youtube.com/watch?v=dXhoTrU1Kkw.

[85] Yoshua Bengio, "Managing AI Risks in an Era of Rapid Progress", November 12, 2023, https://managing-ai-risks.com/.

[86] See:https://www.theguardian.com/technology/2023/oct/24/ai-risk-climate-crisis-google-deepmind-chief-demis-hassabis-regulation.

[87] See https://fortune.com/2023/07/10/anthropic-ceo-dario-amodei-ai-risks-short-medium-long-term/.

[88] For a longer list of leading experts expressing concerns, the reader is asked to visit the stop.ai website (https://www.stop.ai/quotes).

[89] Iason Gabriel, "Artificial intelligence, values, and alignment," *Minds and machines* 30, no. 3 (2020): 411–37; Brian Christian, *The alignment problem: Machine learning and human values* (New York: W.W. Norton & Company, 2020).

[90] Nick Bostrom, "The control problem. Excerpts from superintelligence: Paths, dangers, strategies," *Science Fiction and Philosophy: From Time Travel to Superintelligence* 1, no. 1 (2016): 308–30.

This example is oversimplified and any conscientious reader will presumably immediately come up with many ideas on how to prevent this scenario from being realized, such as asking the agent to merely produce a certain number of paperclips (a goal satisficer, vide the work of Purkitt),[91] adding an off switch or caging the AI agent inside the factory. However, there exist many counter-intuitive theoretical reasons why such safety measures are likely to fail when faced with an AI agent that is sufficiently more intelligent than us.[92]

The emergence of instrumental sub-goals is a particularly interesting phenomenon. It appears that regardless of the specified goal, an agent is incentivized to pursue certain universal sub-goals, such as developing an accurate model of the environment[93] to be able to make better predictions, seek resources such as money, materials, or energy, escape any confinements, and prevent its own shutdown. Conversely, if an agent has reason to anticipate other agents attempting to reduce its power, then it has the incentive to disempower or deceive them first.[94]

Given the previously described and empirically demonstrated risks,[95] it is important that the development of AI is balanced and preceded with research and governance efforts to reduce global risks inherent within AI technologies—so-called differential technology development.[96] With the development of superintelligent agents, it is important that their design incentivizes aligned and secure goal completion; otherwise, we cannot ensure the safety of humanity, as hinted at by Norman Wiener.

In 2022, an expert survey among AI researchers[97] gave a median expectation of 5-10% for the possibility of human extinction due to artificial intelligence. In 2023, after the release of GPT-4, hundreds of leading experts in AI signed an open letter to halt experiments on LLMs for 6 months.[98] Since the letter's release, governments across the world have taken action to pass legislation on this new technology, although their efforts may not be enough to avoid the largest risks.

[91] Helen E. Purkitt, "Artificial intelligence and intuitive foreign policy decision-makers viewed as limited information processors: Some conceptual issues and practical concerns for the future," in *Artificial Intelligence and International Politics* (London: Routledge, 2019), 35–55.

[92] Mark Ring and Laurent Orseau, "Delusion, survival, and intelligent agents," *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011, Proceedings 4* (Springer, 2011), 11–20; Roman Yampolskiy and Joshua Fox, "Safety engineering for artificial general intelligence," *Topoi* 32 (2013): 217–26; Joseph Carlsmith, "Is Power-Seeking AI an Existential Risk?," *arXiv* preprint arXiv:2206.13353 1, no. 1 (2022): 1.

[93] Kenneth Li et al., "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," *The Eleventh International Conference on Learning Representations* (2022).

[94] Tsvi Benson-Tilsen and Nate Soares, "Formalizing Convergent Instrumental Goals," *AAAI Workshop: AI, Ethics, and Society* (2016).

[95] Pan, Bhatia, and Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models," International Conference on Learning Representations (2022).

[96] Jonas Sandbrink et al., "Differential technology development: A responsible innovation principle for navigating technology risks," *SSRN* 1, no. 1 (2022): 4–8.

[97] See https://ourworldindata.org/ai-timelines.

[98] See https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

4. Epilogue

In this position paper, we have outlined the potential risks of existing AI methods to various aspects of our fundamental right to privacy. These include unintentionally leaking sensitive information that can only be inferred by sufficiently powerful AI models with access to large amounts of data, the emerging technology allowing for precise reconstruction of thoughts from non-invasive brain imaging techniques, accurate reconstruction of the floor plans and 3D maps of our homes from WiFi signals, and many more.

We have discussed the potential impacts of AI on the social fabric of our communities through personalized LLM-based chatbots, on the mass-surveillance landscape through facial recognition systems, and even on digital encryption systems through reinforcement learning agents trained to optimize fundamental computer science algorithms such as sorting. We have also briefly outlined the rationale behind existential risks posed by modern AI.

This article has not focused on the many prospective benefits of AI technologies, assuming that they are currently receiving ample attention in existing discourse. Many benefits of AI have already materialized, including protein-folding prediction for the purposes of scientific discovery,[99] cancer screening methods in a medical diagnosis setting that surpass human experts,[100] and even factory work automation.[101] Many more are imminently present on the horizon.

A number of potential technical solutions to remedy some of the risks inherent to AI have been briefly mentioned—in particular, the emerging AI interpretability and safety methods, such as scalable oversight. Readers interested in these methods can find relevant information in the recent works of Bowman and others,[102] Burns and others,[103] and Conmy and others.[104] Another important category of efforts towards increasing the safety of AI methods comes in the form of recent policy developments[105] and global coordination efforts that may help mitigate the stated risks.[106]

It is the hope of these authors that a respectful societal conversation about the risks inherent to wide-scale deployment of AI agents will coincide with further technical develop-

[99] John Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature* 596, no. 7873 (2021): 583–89.

[100] Scott Mayer McKinney et al., "International evaluation of an AI system for breast cancer screening," *Nature* 577, no. 7788 (2020): 89–94.

[101] Dennise Mathew, N.C. Brintha, and J.T. Winowlin Jappes, "Artificial intelligence powered automation for industry 4.0," *New Horizons for Industry 4.0 in Modern Business* (Springer, 2023), 1–28.

[102] Samuel R. Bowman et al., "Measuring progress on scalable oversight for large language models," *arXiv* preprint arXiv:2211.03540 1, no. 1 (2022): 1.

[103] Collin Burns et al., "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision," *arXiv* e-prints, 2023, arXiv–2312.

[104] Arthur Conmy et al., "Towards automated circuit discovery for mechanistic interpretability," *Advances in Neural Information Processing Systems* 36 (2024): 16318-16352.

[105] See https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

[106] Currently, this is primarily in the form of summits such as the AI for Good Global Summit. See https://aiforgood.itu.int/summit24/.

ment of AI inspection and risk mitigation tools as well as concerted, global policy efforts to ensure that any sufficiently powerful AI will be deployed in alignment with our societies' fundamental rights to privacy and more. By shifting our attention to the effects that AI can have on multiple heuristic zones, we can recenter the actual consequences to human life and how the boundaries of privacy can be irreparably invaded.

Disclosure Statement

To the best of their knowledge, the author(s) have no conflicts of interest pertinent to the contents of this article.

Acknowledgements

The authors would like to further thank Matthew Farrugia-Roberts for providing insightful comments regarding the technical aspects of the introduction section of this paper.

References

Altman, Sam. “Planning for AGI and beyond.” OpenAI Blog 1, no. 1 (2023): 1. Accessed 24 February 2023. https://openai.com/blog/planning-for-agi-and-beyond.

Amodei, Dario, Paul Christiano, and Alex Ray. “Learning from human preferences.” 2017. https://openai.com/research/learning-from-human-preferences.

Arulkumaran, Kai, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. “Deep reinforcement learning: A brief survey.” IEEE Signal Processing Magazine 34, no. 6 (2017): 26–38.

Ayers, John W, Zechariah Zhu, Adam Poliak, Eric C. Leas, Mark Dredze, Michael Hogarth, and Davey M Smith. “Evaluating Artificial Intelligence Responses to Public Health Questions.” JAMA Network Open 6, no. 6 (2023): 2317517.

Bae, Gwangbin, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. “Digiface-1m: 1 million digital face images for face recognition.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 3526–35. 2023.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. “Training a helpful and harmless assistant with reinforcement learning from human feedback.” arXiv preprint arXiv:2204.05862 1 (2022): 1.

Baker, Ingmar, Bowen an Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. “Emergent tool use from multi-agent interaction.” 2019. https://openai.com/research/emergent-tool-use.

Bakhtin, Anton, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. “Human-level play in the game of Diplomacy by combining language models with strategic reasoning.” Science 378, no. 6624 (2022): 1067–74.

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. An introduction to ethics in robotics and AI. Springer Nature, 2021.

Benchetrit, Yohann, Hubert Banville, and Jean-Rémi King. “Brain decoding: toward real-time reconstruction of visual perception.” arXiv preprint arXiv:2310.19812 1, no. 1 (2023): 1–10.

Bengio, Yoshua. “Managing AI Risks in an Era of Rapid Progress.” 2023. https://managing-ai-risks.com/.

Benson-Tilsen, Tsvi, and Nate Soares. “Formalizing Convergent Instrumental Goals.” AAAI Workshop: AI, Ethics, and Society. 2016.

Bittle, Jake. “Lie detectors have always been suspect. AI has made the problem worse.” MIT Technology Review, 2020. https://www.technologyreview.com/2020/03/13/905323/ai-lie-detectors-polygraph-silent-talker-iborderctrl-converus-neuroid/.

Bostrom, Nick. “The control problem. Excerpts from superintelligence: Paths, dangers, strategies.” Science Fiction and Philosophy: From Time Travel to Superintelligence 1, no. 1 (2016): 308–30.

Bowman, Samuel R, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukošlute, Amanda Askell, Andy Jones, Anna Chen, et al. “Measuring progress on scalable oversight for large language models.” arXiv preprint arXiv:2211.03540 1, no. 1 (2022): 1.

Branwen, Gwern, Catherine Olsson, Joel Lehman, and Alex Irpa. "Specification gaming examples in AI-master list." 2022. https://heystacks.com/doc/186/specification-gaming-examples-in-ai---master-list.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877–1901.

Bruun, Mette Birkedal. "Towards an Approach to Early Modern Privacy: The Retirement of the Great Condé." In Early Modern Privacy: Sources and Approaches, ed. M. Green, L.C. Nørgaard, and M.B. Bruun, 12–60. Leiden: Brill, 2021. https://doi.org/10.1163/9789004153073_003.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. "Sparks of artificial general intelligence: Early experiments with gpt-4." arXiv preprint arXiv:2303.12712 1 (2023): 1.

Bucknall, Benjamin S., and Shiri Dori-Hacohen. "Current and near-term AI as a potential existential risk factor." Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 119–129. 2022.

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision." arXiv e-prints, 2023, arXiv–2312.

Burns, John Lee, Zachary Zaiman, Jack Vanschaik, Gaoxiang Luo, Le Peng, Brandon Price, Garric Mathias, Vijay Mittal, Akshay Sagane, Christopher Tignanelli, et al. "Ability of artificial intelligence to identify self-reported race in chest x-ray using pixel intensity counts." Journal of Medical Imaging 10, no. 6 (2023): 61106.

Carlsmith, Joseph. "Is Power-Seeking AI an Existential Risk?" arXiv preprint arXiv:2206.13353 1, no. 1 (2022): 1.

Cetinic, Eva, and James She. "Understanding and creating art with AI: Review and outlook." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18, no. 2 (2022): 1–22.

Christian, Brian. The alignment problem: Machine learning and human values. W.W. Norton & Company, 2020.

Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep reinforcement learning from human preferences." Advances in neural information processing systems 30 (2017): 1–10.

Clark, Jack, and Dario Amodei. "Faulty reward functions in the wild." 2016. https://openai.com/research/faulty-reward-functions.

Conmy, Arthur, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. "Towards automated circuit discovery for mechanistic interpretability." Advances in Neural Information Processing Systems 36 (2024): 16318-16352.

Cunningham, Pádraig, Matthieu Cord, and Sarah Jane Delany. "Supervised learning." In Machine learning techniques for multimedia: case studies on organization and retrieval, 21–49. Springer, 2008.

Cybenko, George. "Approximation by superpositions of a sigmoidal function." Mathematics of control, signals and systems 2, no. 4 (1989): 303–14.

D'Alfonso, Simon. "AI in mental health." Current Opinion in Psychology 36 (2020): 112–17.

Damian, Alexandru, Jason Lee, and Mahdi Soltanolkotabi. "Neural networks can learn representations with gradient descent." In Conference on Learning Theory, 5413–52. PMLR, 2022.

Dokmanic, Ivan, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. "Acoustic echoes reveal room shape." Proceedings of the National Academy of Sciences 110, no. 30 (2013): 12186–91.

Dong, Zizhao, Gang Wang, Shaoyuan Lu, Luyao Dai, Shucheng Huang, and Ye Liu. "Intentional-deception detection based on facial muscle movements in an interactive social context." Pattern Recognition Letters 164 (2022): 30–9.

Fiske, Amelia, Peter Henningsen, and Alena Buyx. "Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy." Journal of medical Internet research 21, no. 5 (2019): 13216.

Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." Minds and Machines 30 (2020): 681–94.

Gabriel, Iason. "Artificial intelligence, values, and alignment." Minds and machines 30, no. 3 (2020): 411–37.

Ganguli, Deep, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. "Predictability and surprise in large generative models." In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (2022), 1747–64.

Gao, Ruohan, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. "Visualechoes: Spatial image representation learning through echolocation." Computer Vision–ECCV 2020. 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, 658–76. Springer, 2020.

Garcia-Navarro, Lulu. "Creating A 'Dadbot' To Talk With A Dead Father." 2017. https://www.npr.org/2017/07/23/538825555/creating-a-dadbot-to-talk-with-a-dead-father.

Geng, Jiaqi, Dong Huang, and Fernando De la Torre. "DensePose From WiFi." arXiv e-prints -, nos. - (2022): arXiv–2301.

Gichoya, Judy Wawira, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. "AI recognition of patient race in medical imaging: a modelling study." The Lancet Digital Health 4, no. 6 (2022): 406–14.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. "When will AI exceed human performance? Evidence from AI experts." Journal of Artificial Intelligence Research 62 (2018): 729–54.

Han, Lawrence, and Hao Tang. "Designing of Prompts for Hate Speech Recognition with In-Context Learning." In 2022 International Conference on Computational Science and Computational Intelligence (CSCI), 319–20. IEEE, 2022.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Universal approximation of an unknown mapping and its derivatives using multilayer feed-forward networks." Neural networks 3, no. 5 (1990): 551–60.

Hyde, Steven J., Eric Bachura, Jonathan Bundy, Richard T. Gretz, and Wm Gerard Sanders. "The tangled webs we weave: Examining the effects of CEO deception on analyst recommendations." Strategic Management Journal 45, no. 1 (2024): 66–112.

Jabbour, Sarah, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. "Deep learning applied to chest x-rays: Exploiting and preventing shortcuts." In Machine Learning for Healthcare Conference, 750–82. PMLR, 2020.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596, no. 7873 (2021): 583–9.

Karras, Johanna, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. "Dreampose: Fashion video synthesis with stable diffusion." In Proceedings of the IEEE/CVF International Conference on Computer Vision (2023), 22680–90.

Kosinski, Michal. "Theory of mind might have spontaneously emerged in large language models." arXiv preprint https://arxiv. org/abs/2302.02083 1, no. 1 (2023): 1–2.

Kosinski, Michal, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior." Proceedings of the national academy of sciences 110, no. 15 (2013): 5802–805.

Kratsios, Anastasis, and Ievgen Bilokopytov. "Non-euclidean universal approximation." Advances in Neural Information Processing Systems 33 (2020): 10635–46.

Levchev, Plamen, Michael N. Krishnan, Chaoran Yu, Joseph Menke, and Avideh Zakhor. "Simultaneous fingerprinting and mapping for multimodal image and WiFi indoor positioning." In 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 442–50. IEEE, 2014.

Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." The Eleventh International Conference on Learning Representations (2022).

Lin, Jinying, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. "A review on interactive reinforcement learning from human social feedback." IEEE Access 8 (2020): 120757–65.

Liu, Kan, and Lu Chen. "Deep Neural Network Learning for Medical Triage." Data Analysis and Knowledge Discovery 3, no. 6 (2019): 99–108.

Lovens, Pierre-François. "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là." (2023). https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC-5WRDX7J2RCHNWPDST24/.

Lu, Yulong, and Jianfeng Lu. "A universal approximation theorem of deep neural networks for expressing probability distributions." Advances in neural information processing systems 33 (2020): 3094–105.

Lu, Zhou, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. "The expressive power of neural networks: A view from the width." Advances in neural information processing systems 30 (2017): 26–38.

Maddocks, Sophie. "'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes." Porn Studies 7, no. 4 (2020): 415–23.

Mankowitz, Daniel J, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, et al.

"Faster sorting algorithms discovered using deep reinforcement learning." Nature 618, no. 7964 (2023): 257–63.

Masood, Momina, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward." Applied intelligence 53, no. 4 (2023): 3974–4026.

Mathew, Dennise, N.C. Brintha, and J.T. Winowlin Jappes. "Artificial intelligence powered automation for industry 4.0." In New Horizons for Industry 4.0 in Modern Business, 1–28. Springer, 2023.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, et al. "International evaluation of an AI system for breast cancer screening." Nature 577, no. 7788 (2020): 89–94.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks." Digital signal processing 73 (2018): 1–15.

Nanda, Neel, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. "Progress measures for grokking via mechanistic interpretability." arXiv preprint arXiv:2301.05217 1, no. 1 (2023): 1–12.

Niikura, Ryota, Tomonori Aoki, Satoki Shichijo, Atsuo Yamada, Takuya Kawahara, Yusuke Kato, Yoshihiro Hirata, Yoku Hayakawa, Nobumi Suzuki, Masanori Ochi, et al. "Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy." Endoscopy 54, no. 8 (2022): 780–4.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. "Bias in data-driven artificial intelligence systems—An introductory survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, no. 3 (2020): 1356.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt. "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models." International Conference on Learning Representations. 2022.

Pan, Yue, and Limao Zhang. "Roles of artificial intelligence in construction engineering and management: A critical review and future trends." Automation in Construction 122 (2021): 103517.

Purkitt, Helen E. "Artificial intelligence and intuitive foreign policy decision-makers viewed as limited information processors: Some conceptual issues and practical concerns for the future." In Artificial Intelligence and International Politics, 35–55. London: Routledge, 2019.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. "Language models are unsupervised multitask learners." OpenAI Blog 1, no. 8 (2019): 9.

Ring, Mark, and Laurent Orseau. "Delusion, survival, and intelligent agents." Artificial General Intelligence. 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings 4, 11–20. Springer, 2011.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022), 10684–95.

Russell, Stuart. "Artificial Intelligence: A Modern Approach, 2021. Human-compatible artificial intelligence." Human-like machine intelligence 1, no. 1 (2021): 3–23.

Sadiku, Matthew N.O., Tolulope J. Ashaolu, Abayomi Ajayi-Majebi, and Sarhan M. Musa. "Artificial intelligence in social media." International Journal of Scientific Advances 2, no. 1 (2021): 15–20.

Sandbrink, Jonas, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg. "Differential technology development: A responsible innovation principle for navigating technology risks." Available at SSRN 1, no. 1 (2022): 4–8.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision (2017), 618–26.

Shinn, Noah, Beck Labash, and Ashwin Gopinath. "Reflexion: an autonomous agent with dynamic memory and self-reflection." arXiv preprint arXiv:2303.11366 1, no. 1 (2023): 1–10.

Shuster, Anastasia, Lilah Inzelberg, Ori Ossmy, Liz Izakson, Yael Hanein, and Dino J. Levy. "Lie to my face: An electromyography approach to the study of deceptive behavior." Brain and Behavior 11, no. 12 (2021): 2386.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. "Mastering the game of Go with deep neural networks and tree search." Nature 529, no. 7587 (2016): 484–89.

Strachan, James, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Alessandro Rufo, Guido Manzi, Michael Graziano, and Cristina Becchio. "Testing Theory of Mind in GPT Models and Humans." arXiv preprint 1, no. 1 (2023): 1–13.

Sun, Chenbo. "Comparative Study of RSA Encryption and Quantum Encryption." Theoretical and Natural Science 2, no. 1 (2023): 121–25.

Sun, Penghao, Zehua Guo, Sen Liu, Julong Lan, Junchao Wang, and Yuxiang Hu. "Smart-FCT: Improving power-efficiency for data center networks with deep reinforcement learning." Computer Networks 179 (2020): 107255.

Szymanski, Nathan J., Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. "An autonomous laboratory for the accelerated synthesis of novel materials." Nature 1, no. 1 (2023): 1–6.

Tang, Jerry, Amanda LeBel, Shailee Jain, and Alexander G. Huth. "Semantic reconstruction of continuous language from non-invasive brain recordings." Nature Neuroscience 1, no. 1 (2023): 1–9.

Turing, Sara. Alan M. Turing: Centenary Edition. Cambridge: Cambridge University Press, 2012.

Vigliotti, Jonathan. "How AI is transforming Hollywood, and why it's at the center of contract negotiations." 2023. https://www.cbsnews.com/news/artificial-intelligence-actors-strike-sag-aftra-metaphysic/.

Vinyals, Oriol, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning." Nature 575, no. 7782 (2019): 350–54.

Wang, Ben (@kingoflolz), and Aran Komatsuzaki. "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model," May 15, 2022. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, Zuoguang, Hongsong Zhu, and Limin Sun. "Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods." IEEE Access 9 (2021): 11895–910.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. "Taxonomy of risks posed by language models." In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (2022). 214–29.

Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." Communications of the ACM 9, no. 1 (1966): 36–45.

Westerlund, Mika. "The emergence of deepfake technology: A review." Technology innovation management review 9, no. 11 (2019): 1–10.

Wiener, Norbert. "Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers." Science 131, no. 3410 (1960): 1355–58.

Yampolskiy, Roman, and Joshua Fox. "Safety engineering for artificial general intelligence." Topoi 32 (2013): 217–26.

Zimmermann, Roland S., Thomas Klein, and Wieland Brendel. "Scale alone does not improve mechanistic interpretability in vision models." arXiv preprint arXiv:2307.05471 1, no. 1 (2023): 1.