RESEARCH COMPILATION

Volume 1

Topic: "AI Safety"

Generated: 05/02/2026, 17:28:02
Total Sources: 2

TABLE OF CONTENTS

AI safety

Type: Ø<ß Web Page
URL: https://en.wikipedia.org/wiki/AI_safety
Date Extracted: 2026-02-05T17:28:01.978Z
Word Count: 9,419 words

AI safety is an interdisciplinary field focused on preventing accidents, misuse, or other harmful consequences arising from artificial intelligence (AI) systems. It encompasses AI alignment (which aims to ensure AI systems behave as intended), monitoring AI systems for risks, and enhancing their robustness. The field is particularly concerned with existential risks posed by advanced AI models.[1][2]

Beyond technical research, AI safety involves developing norms and policies that promote safety. It gained significant popularity in 2023, with rapid progress in generative AI and public concerns voiced by researchers and CEOs about potential dangers. During the 2023 AI Safety Summit, the United States and the United Kingdom both established their own AI Safety Institute. However, researchers have expressed concern that AI safety measures are not keeping pace with the rapid development of AI capabilities.[3] Scholars discuss current risks from critical systems failures,[4] bias,[5] and AI-enabled surveillance,[6] as well as emerging risks like technological unemployment, digital manipulation,[7] weaponization,[8] AI-enabled cyberattacks[9] and bioterrorism.[10] They also discuss speculative risks from losing control of future artificial general intelligence (AGI) agents,[11] or from AI enabling perpetually stable dictatorships.[12] Some have criticized concerns about AGI, such as Andrew Ng who compared them in 2015 to "worrying about overpopulation on Mars when we have not even set foot on the planet yet".[13] Stuart J. Russell on the other side urges caution, arguing that "it is better to anticipate human ingenuity than to underestimate it".[14]

AI researchers have widely differing opinions about the severity and primary sources of risk posed by AI technology[15][16][17] – though surveys suggest that experts take high consequence risks seriously. In two surveys of AI researchers, the median respondent was optimistic about AI overall, but placed a 5% probability on an "extremely bad (e.g. human extinction)" outcome of advanced AI.[15] In a 2022 survey of the natural language processing community, 37% agreed or weakly agreed that it is plausible that AI decisions could lead to a catastrophe that is "at least as bad as an all-out nuclear war".[18] Risks from AI began to be seriously discussed at the start of the computer age: Moreover, if we move in the direction of making machines which learn and whose behavior is modified by experience, we must face the fact that every degree of independence we give the machine is a degree of possible defiance of our wishes.

In 1988 Blay Whitby published a book outlining the need for AI to be developed along ethical and socially responsible lines.[20] From 2008 to 2009, the Association for the Advancement of Artificial Intelligence (AAAI) commissioned a study to explore and address potential long-term societal influences of AI research and development. The panel was generally skeptical of the radical views expressed by science-fiction authors but agreed that "additional research would be valuable on methods for understanding and verifying the range of behaviors of complex computational systems to minimize unexpected outcomes".[21]

In 2011, Roman Yampolskiy introduced the term "AI safety engineering"[22] at the Philosophy and Theory of Artificial Intelligence conference,[23] listing prior failures of AI systems and arguing that "the frequency and seriousness of such events will steadily increase as AIs become more capable".[24]

In 2014, philosopher Nick Bostrom published the book Superintelligence: Paths, Dangers, Strategies. He has the opinion that the rise of AGI has the potential to create various societal issues, ranging from the displacement of the workforce by AI, manipulation of political and military structures, to even the possibility of human extinction.[25] His argument that future advanced systems may pose a threat to human existence prompted Elon Musk,[26] Bill Gates,[27] and Stephen Hawking[28] to voice similar concerns.

In 2015, dozens of artificial intelligence experts signed an open letter on artificial intelligence calling for research on the societal impacts of AI and outlining concrete directions.[29] To date, the letter has been signed by over 8000 people including Yann LeCun, Shane Legg, Yoshua Bengio, and Stuart Russell.

In the same year, a group of academics led by professor Stuart J. Russell founded the Center for Human-Compatible AI at the University of California Berkeley and the Future of Life Institute awarded $6.5 million in grants for research aimed at "ensuring artificial intelligence (AI) remains safe, ethical and beneficial".[30]

In 2016, the White House Office of Science and Technology Policy and Carnegie Mellon University announced The Public Workshop on Safety and Control for Artificial Intelligence,[31] which was one of a sequence of four White House workshops aimed at investigating "the advantages and drawbacks" of AI.[32] In the same year, Concrete Problems in AI Safety – one of the first and most influential technical AI Safety agendas – was published.[33]

In 2017, the Future of Life Institute sponsored the Asilomar Conference on Beneficial AI, where more than 100 thought leaders formulated principles for beneficial AI including "Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards".[34]

In 2018, the DeepMind Safety team outlined AI safety problems in specification, robustness,[35] and assurance.[36] The following year, researchers organized a workshop at ICLR that focused on these problem areas.[37]

In 2021, Unsolved Problems in ML Safety was published, outlining research directions in robustness, monitoring, alignment, and systemic safety.[2]

In 2023, Rishi Sunak said he wants the United Kingdom to be the "geographical home of global AI safety regulation" and to host the first global summit on AI safety.[38] The AI safety summit took place in November 2023, and focused on the risks of misuse and loss of control associated with frontier AI models.[39] During the summit the intention to create the International Scientific Report on the Safety of Advanced AI[40] was announced.

In 2024, The US and UK forged a new partnership on the science of AI safety. The MoU was signed on 1 April 2024 by US commerce secretary Gina Raimondo and UK technology secretary Michelle Donelan to jointly develop advanced AI model testing, following commitments announced at an AI Safety Summit in Bletchley Park in November.[41]

In 2025, an international team of 96 experts chaired by Yoshua Bengio published the first International AI Safety Report. The report, commissioned by 30 nations and the United Nations, represents the first global scientific review of potential risks associated with advanced artificial intelligence. It details potential threats stemming from misuse, malfunction, and societal disruption, with the objective of informing policy through evidence-based findings, without providing specific recommendations.[42][43] AI safety research areas include robustness, monitoring, and alignment.[2][36] Adversarial robustness[edit]

AI systems are often vulnerable to adversarial examples or "inputs to machine learning (ML) models that an attacker has intentionally designed to cause the model to make a mistake".[44] For example, in 2013, Szegedy et al. discovered that adding specific imperceptible perturbations to an image could cause it to be misclassified with high confidence.[45] This continues to be an issue with neural networks, though in recent work the perturbations are generally large enough to be perceptible.[46][47][48] Carefully crafted noise can be added to an image to cause it to be misclassified with high confidence.

The image on the right is predicted to be an ostrich after the perturbation is applied. (Left) is a correctly predicted sample, (center) perturbation applied magnified by 10x, (right) adversarial example.[45]

Adversarial robustness is often associated with security.[49] Researchers demonstrated that an audio signal could be imperceptibly modified so that speech-to-text systems transcribe it to any message the attacker chooses.[50] Network intrusion[51] and malware[52] detection systems also must be adversarially robust since attackers may design their attacks to fool detectors.

Models that represent objectives (reward models) must also be adversarially robust. For example, a reward model might estimate how helpful a text response is and a language model might be trained to maximize this score.[53] Researchers have shown that if a language model is trained for long enough, it will leverage the vulnerabilities of the reward model to achieve a better score and perform worse on the intended task.[54] This issue can be addressed by improving the adversarial robustness of the reward model.[55] More generally, any AI system used to evaluate another AI system must be adversarially robust. This could include monitoring tools, since they could also potentially be tampered with to produce a higher reward.[56]

Large language models (LLMs) can be vulnerable to prompt injection[57] and model stealing,[58] and may be used to generate misinformation.[59] Prompt injection involves embedding instructions into prompts in order to bypass safety measures.[57] Estimating uncertainty[edit]

It is often important for human operators to gauge how much they should trust an AI system, especially in high-stakes settings such as medical diagnosis.[60] ML models generally express confidence by outputting probabilities; however, they are often overconfident,[61] especially in situations that differ from those that they were trained to handle.[62] Calibration research aims to make model probabilities correspond as closely as possible to the true proportion that the model is correct.

Similarly, anomaly detection or out-of-distribution (OOD) detection aims to identify when an AI system is in an unusual situation. For example, if a sensor on an autonomous vehicle is malfunctioning, or it encounters challenging terrain, it should alert the driver to take control or pull over.[63] Anomaly detection has been implemented by simply training a classifier to distinguish anomalous and non-anomalous inputs,[64] though a range of additional techniques are in use.[65][66] Detecting malicious use[edit]

Scholars[8] and government agencies have expressed concerns that AI systems could be used to help malicious actors to build weapons,[67] manipulate public opinion,[68][69] or automate cyber attacks.[70] These worries are a practical concern for companies like OpenAI which host powerful AI tools online.[71] In order to prevent misuse, OpenAI has built detection systems that flag or restrict users based on their activity.[72] Neural networks have often been described as black boxes,[73] meaning that it is difficult to understand why they make the decisions they do as a result of the massive number of computations they perform.[74] This makes it challenging to anticipate failures. In 2018, a self-driving car killed a pedestrian after failing to identify them. Due to the black box nature of the AI software, the reason for the failure remains unclear.[75] It also raises debates in healthcare over whether statistically efficient but opaque models should be used.[76]

One critical benefit of transparency is explainability.[77] It is sometimes a legal requirement to provide an explanation for why a decision was made in order to ensure fairness, for example for automatically filtering job applications or credit score assignment.[77]

Another benefit is to reveal the cause of failures.[73] At the beginning of the 2020 COVID-19 pandemic, researchers used transparency tools to show that medical image classifiers were 'paying attention' to irrelevant hospital labels.[78]

Transparency techniques can also be used to correct errors. For example, in the paper "Locating and Editing Factual Associations in GPT", the authors were able to identify model parameters that influenced how it answered questions about the location of the Eiffel tower. They were then able to 'edit' this knowledge to make the model respond to questions as if it believed the tower was in Rome instead of France.[79] Though in this case, the authors induced an error, these methods could potentially be used to efficiently fix them. Model editing techniques also exist in computer vision.[80]

Finally, some have argued that the opaqueness of AI systems is a significant source of risk and better understanding of how they function could prevent high-consequence failures in the future.[81] "Inner" interpretability research aims to make ML models less opaque. One goal of this research is to identify what the internal neuron activations represent.[82][83] For example, researchers identified a neuron in the CLIP artificial intelligence system that responds to images of people in Spider-Man costumes, sketches of Spider-Man, and the word 'spider'.[84] It also involves explaining connections between these neurons or 'circuits'.[85][86] For example, researchers have identified pattern-matching mechanisms in transformer attention that may play a role in how language models learn from their context.[87] "Inner interpretability" has been compared to neuroscience. In both cases, the goal is to understand what is going on in an intricate system, though ML researchers have the benefit of being able to take perfect measurements and perform arbitrary ablations.[88] Machine learning models can potentially contain "trojans" or "backdoors": vulnerabilities that bad actors maliciously build into an AI system. For example, a trojaned facial recognition system could grant access when a specific piece of jewelry is in view;[2] or a trojaned autonomous vehicle may function normally until a specific trigger is visible.[89] This might not be difficult to do with some large models like CLIP or GPT-3 as they are trained on publicly available internet data.[90] Researchers were able to plant a trojan in an image classifier by changing just 300 out of 3 million of the training images.[91] In addition to posing a security risk, researchers have argued that trojans provide a concrete setting for testing and developing better monitoring tools.[56]

A 2024 research paper by Anthropic showed that large language models could be trained with persistent backdoors. These "sleeper agent" models could be programmed to generate malicious outputs (such as vulnerable code) after a specific date, while behaving normally beforehand. Standard AI safety measures, such as supervised fine-tuning, reinforcement learning and adversarial training, failed to remove these backdoors.[92] In the field of artificial intelligence (AI), alignment aims to steer AI systems toward a person's or group's intended goals, preferences, or ethical principles. An AI system is considered aligned if it advances the intended objectives. A misaligned AI system pursues unintended objectives.[93]

It is often challenging for AI designers to align an AI system because it is difficult for them to specify the full range of desired and undesired behaviors. Therefore, AI designers often use simpler proxy goals, such as gaining human approval. But proxy goals can overlook necessary constraints or reward the AI system for merely appearing aligned.[93][94] AI systems may also find loopholes that allow them to accomplish their proxy goals efficiently but in unintended, sometimes harmful, ways (reward hacking).[93][95]

Advanced AI systems may develop unwanted instrumental strategies, such as seeking power or survival because such strategies help them achieve their assigned final goals.[93][96][97] Furthermore, they might develop undesirable emergent goals that could be hard to detect before the system is deployed and encounters new situations and data distributions.[98][99] Empirical research showed in 2024 that advanced large language models (LLMs) such as OpenAI o1 or Claude 3 sometimes engage in strategic deception to achieve their goals or prevent them from being changed.[100][101]

Today, some of these issues affect existing commercial systems such as LLMs,[102][103][104] robots,[105] autonomous vehicles,[106] and social media recommendation engines.[102][97][107] Some AI researchers argue that more capable future systems will be more severely affected because these problems partially result from high capabilities.[108][95][94]

Many prominent AI researchers and the leadership of major AI companies have argued or asserted that AI is approaching human-like (AGI) and superhuman cognitive capabilities (ASI), and could endanger human civilization if misaligned.[109][97] These include "AI godfathers" Geoffrey Hinton and Yoshua Bengio and the CEOs of OpenAI, Anthropic, and Google DeepMind.[110][111][112] These risks remain debated.[113]

[Continued - Source 1: AI safety...]

AI alignment is a subfield of AI safety, the study of how to build safe AI systems.[114][115] Other subfields of AI safety include robustness, monitoring, and capability control.[116] Research challenges in alignment include instilling complex values in AI, developing honest AI, scalable oversight, auditing and interpreting AI models, and preventing emergent AI behaviors like power-seeking.[116] Alignment research has connections to interpretability research,[117][118] (adversarial) robustness,[119] anomaly detection, calibrated uncertainty, [117] formal verification,[120] preference learning,[121][122][123] safety-critical engineering,[124] game theory,[125] algorithmic fairness,[119][126] and social sciences.[127][128] Systemic safety and sociotechnical factors[edit]

It is common for AI risks (and technological risks more generally) to be categorized as misuse or accidents. [129] Some scholars have suggested that this framework falls short.[129] For example, the Cuban Missile Crisis was not clearly an accident or a misuse of technology.[129] Policy analysts Zwetsloot and Dafoe wrote, "The misuse and accident perspectives tend to focus only on the last step in a causal chain leading up to a harm: that is, the person who misused the technology, or the system that behaved in unintended ways... Often, though, the relevant causal chain is much longer." Risks often arise from 'structural' or 'systemic' factors such as competitive pressures, diffusion of harms, fast-paced development, high levels of uncertainty, and inadequate safety culture.[129] In the broader context of safety engineering, structural factors like 'organizational safety culture' play a central role in the popular STAMP risk analysis framework.[130]

Inspired by the structural perspective, some researchers have emphasized the importance of using machine learning to improve sociotechnical safety factors, for example, using ML for cyber defense, improving institutional decision-making, and facilitating cooperation.[2] Others have emphasized the importance of involving both AI practitioners and domain experts in the design process to address structural vulnerabilities. [131] Some scholars are concerned that AI will exacerbate the already imbalanced game between cyber attackers and cyber defenders.[132] This would increase 'first strike' incentives and could lead to more aggressive and destabilizing attacks. In order to mitigate this risk, some have advocated for an increased emphasis on cyber defense. In addition, software security is essential for preventing powerful AI models from being stolen and misused.[8] Recent studies have shown that AI can significantly enhance both technical and managerial cybersecurity tasks by automating routine tasks and improving overall efficiency.[133] AI safety research has also examined defensive techniques for protecting machine learning systems from data poisoning attacks during training. In particular, label-flipping attacks can degrade model performance while remaining difficult to detect using conventional data validation methods. To address this risk, recent work has proposed model-agnostic detection pipelines that monitor learning behaviour and combine multiple detectors to identify suspicious training samples. Such approaches aim to strengthen cyber defense by improving the resilience and trustworthiness of AI systems operating in adversarial settings.[134][135] Improving institutional decision-making[edit]

The advancement of AI in economic and military domains could precipitate unprecedented political challenges. [136] Some scholars have compared AI race dynamics to the cold war, where the careful judgment of a small number of decision-makers often spelled the difference between stability and catastrophe.[137] AI researchers have argued that AI technologies could also be used to assist decision-making.[2] For example, researchers are beginning to develop AI forecasting[138] and advisory systems.[139] Facilitating cooperation[edit]

Many of the largest global threats (nuclear war,[140] climate change,[141] etc.) have been framed as cooperation challenges. As in the well-known prisoner's dilemma scenario, some dynamics may lead to poor results for all players, even when they are optimally acting in their self-interest. For example, no single actor has strong incentives to address climate change even though the consequences may be significant if no one intervenes.[141]

A salient AI cooperation challenge is avoiding a 'race to the bottom'.[142] In this scenario, countries or companies race to build more capable AI systems and neglect safety, leading to a catastrophic accident that harms everyone involved. Concerns about scenarios like these have inspired both political[143] and technical[144] efforts to facilitate cooperation between humans, and potentially also between AI systems. Most AI research focuses on designing individual agents to serve isolated functions (often in 'single-player' games).[145] Scholars have suggested that as AI systems become more autonomous, it may become essential to study and shape the way they interact.[145][131] The AI Safety Summit of November 2023[146]

AI governance is broadly concerned with creating norms, standards, and regulations to guide the use and development of AI systems.[137] In AI safety, local solutions focus on individual AI systems, ensuring they are safe and beneficial, while global solutions seek to implement safety measures for all AI systems across various jurisdictions.[147]

AI safety governance research ranges from foundational investigations into the potential impacts of AI to specific applications. On the foundational side, researchers have argued that AI could transform many aspects of society due to its broad applicability, comparing it to electricity and the steam engine.[148] Some work has focused on anticipating specific risks that may arise from these impacts – for example, risks from mass unemployment,[149] weaponization,[150] disinformation,[151] surveillance,[152] and the concentration of power.[153] Other work explores underlying risk factors such as the difficulty of monitoring the rapidly evolving AI industry,[154] the availability of AI models,[155] and 'race to the bottom' dynamics.[142][156] Allan Dafoe, the head of longterm governance and strategy at DeepMind has emphasized the dangers of racing and the potential need for cooperation: "it may be close to a necessary and sufficient condition for AI safety and alignment that there be a high degree of caution prior to deploying advanced powerful systems; however, if actors are competing in a domain with large returns to first-movers or relative advantage, then they will be pressured to choose a sub-optimal level of caution".[143] A research stream focuses on developing approaches, frameworks, and methods to assess AI accountability, guiding and promoting audits of AI-based systems.[157][158][159] A key challenge for these approaches is a lack of widely accepted standards, and ambiguity about what the methods would require,[160][161] as well as a lack of safety culture in the industry.[162]

Efforts to enhance AI safety include frameworks designed to align AI outputs with ethical guidelines and reduce risks like misuse and data leakage. Tools such as Nvidia's Guardrails,[163] Llama Guard,[164] Preamble's customizable guardrails[165] and Claude's Constitution mitigate vulnerabilities like prompt injection and ensure outputs adhere to predefined principles. These frameworks are often integrated into AI systems to improve safety and reliability.[166] Philosophical perspectives[edit] The field of AI safety is deeply intertwined with philosophical considerations, particularly in the realm of ethics. Deontological ethics, which emphasizes adherence to moral rules, has been proposed as a framework for aligning AI systems with human values. Some have suggested that by embedding deontological principles, AI systems can be guided to avoid actions that cause harm, ensuring their operations remain within ethical boundaries,[167] but those suggestions have been questioned, with other alternatives being suggested at more promising.[168] Some experts have argued that it is too early to regulate AI, expressing concerns that regulations will hamper innovation and it would be foolish to "rush to regulate in ignorance".[169][170] Others, such as business magnate Elon Musk, call for pre-emptive action to mitigate catastrophic risks.[171]

Outside of formal legislation, government agencies have put forward ethical and safety recommendations. In March 2021, the US National Security Commission on Artificial Intelligence reported that advances in AI may make it increasingly important to "assure that systems are aligned with goals and values, including safety, robustness and trustworthiness".[172] Subsequently, the National Institute of Standards and Technology drafted a framework for managing AI Risk, which advises that when "catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed".[173]

In September 2021, the People's Republic of China (PRC) published ethical guidelines for the use of AI in China, emphasizing that AI decisions should remain under human control and calling for accountability mechanisms. In the same month, The United Kingdom published its 10-year National AI Strategy,[174] which states the British government "takes the long-term risk of non-aligned Artificial General Intelligence, and the unforeseeable changes that it would mean for ... the world, seriously".[175] The strategy describes actions to assess long-term AI risks, including catastrophic risks.[175] The British government held first major global summit on AI safety. This took place on the 1st and 2 November 2023 and was described as "an opportunity for policymakers and world leaders to consider the immediate and future risks of AI and how these risks can be mitigated via a globally coordinated approach".[176][177] China Media Project stated "key aspects of its approach remain fundamentally unsafe by the standards of democratic societies worldwide", arguing that part of China's AI safety approach is focused on strengthening the CCP's information control.[178]

Government organizations, particularly in the United States, have also encouraged the development of technical AI safety research. The Intelligence Advanced Research Projects Activity initiated the TrojAI project to identify and protect against Trojan attacks on AI systems.[179] The DARPA engages in research on explainable artificial intelligence and improving robustness against adversarial attacks.[180][181] And the National Science Foundation supports the Center for Trustworthy Machine Learning, and is providing millions of dollars in funding for empirical AI safety research.[182]

In 2024, the United Nations General Assembly adopted the first global resolution on the promotion of "safe, secure and trustworthy" AI systems that emphasized the respect, protection and promotion of human rights in the design, development, deployment and the use of AI.[183]

In May 2024, the Department for Science, Innovation and Technology (DSIT) announced £8.5 million in funding for AI safety research under the Systemic AI Safety Fast Grants Programme, led by Christopher Summerfield and Shahar Avin at the AI Safety Institute, in partnership with UK Research and Innovation. Technology Secretary Michelle Donelan announced the plan at the AI Seoul Summit, stating the goal was to make AI safe across society and that promising proposals could receive further funding. The UK also signed an agreement with 10 other countries and the EU to form an international network of AI safety institutes to promote collaboration and share information and resources. Additionally, the UK AI Safety Institute planned to open an office in San Francisco.[184] Corporate self-regulation[edit]

AI labs and companies generally abide by safety practices and norms that fall outside of formal legislation.[185] One aim of governance researchers is to shape these norms. Examples of safety recommendations found in the literature include performing third-party auditing,[186] offering bounties for finding failures,[186] sharing AI incidents[186] (an AI incident database was created for this purpose),[187] following guidelines to determine whether to publish research or models,[155] and improving information and cyber security in AI labs.[188]

Companies have also made commitments. Cohere, OpenAI, and AI21 proposed and agreed on "best practices for deploying language models", focusing on mitigating misuse.[189] To avoid contributing to racing-dynamics, OpenAI has also stated in their charter that "if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project"[190] Also, industry leaders such as CEO of DeepMind Demis Hassabis, director of Facebook AI Yann LeCun have signed open letters such as the Asilomar Principles[34] and the Autonomous Weapons Open Letter.[191] AI alignment

Artificial intelligence and elections

Artificial intelligence detection software

Hallucination (artificial intelligence) ^ Ahmed, Shazeda; Ja §v‘Dska, Klaudia; Ahlawat, Archana; Winecoff, Amy; Wang, Mona (2024-04-14). "Field-building and the epistemic culture of AI safety". First Monday. doi:10.5210/fm.v29i4.13626. ISSN 1396-0466. ^ a b c d e f Hendrycks, Dan; Carlini, Nicholas; Schulman, John; Steinhardt, Jacob (2022-06-16). "Unsolved Problems in ML Safety". arXiv:2109.13916. ^ Perrigo, Billy (2023-11-02). "U.K.'s AI Safety Summit Ends With Limited, but Meaningful, Progress". Time. Retrieved 2024-06-02. ^ De-Arteaga, Maria (2020-05-13). Machine Learning in High-Stakes Settings: Risks and Opportunities (PhD). Carnegie Mellon University. ^ Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina; Galstyan, Aram (2021). "A Survey on Bias and Fairness in Machine Learning". ACM Computing Surveys. 54 (6): 1–35. arXiv:1908.09635. doi:10.1145/3457607. ISSN 0360-0300. S2CID 201666566. Archived from the original on 2022-11-23. Retrieved 2022-11-28. ^ Feldstein, Steven (2019). The Global Expansion of AI Surveillance (Report). Carnegie Endowment for International Peace. ^ Barnes, Beth (2021). "Risks from AI persuasion". Lesswrong. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ a b c Brundage, Miles; Avin, Shahar; Clark, Jack; Toner, Helen; Eckersley, Peter; Garfinkel, Ben; Dafoe, Allan; Scharre, Paul; Zeitzoff, Thomas; Filar, Bobby; Anderson, Hyrum; Roff, Heather; Allen, Gregory C; Steinhardt, Jacob; Flynn, Carrick (2018-04-30). "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". Apollo-University Of Cambridge Repository, Apollo-University Of Cambridge Repository. Apollo - University of Cambridge Repository. doi:10.17863/cam.22520. S2CID 3385567. Archived from the original on 2022-11-23. Retrieved 2022-11-28. ^ Davies, Pascale (December 26, 2022). "How NATO is preparing for a new era of AI cyber attacks". euronews. Retrieved 2024-03-23. ^ Ahuja, Anjana (February 7, 2024). "AI's bioterrorism potential should not be ruled out". Financial Times. Retrieved 2024-03-23. ^ Carlsmith, Joseph (2022-06-16). "Is Power-Seeking AI an Existential Risk?". arXiv:2206.13353. ^ Minardi, Di (16 October 2020). "The grim fate that could be 'worse than extinction'". BBC. Retrieved 2024-03-23. ^ "AGI Expert Peter Voss Says AI Alignment Problem is Bogus | NextBigFuture.com". 2023-04-04. Retrieved 2023-07-23. ^ Dafoe, Allan (2016). "Yes, We Are Worried About the Existential Risk of Artificial Intelligence". MIT Technology Review. Archived from the original on 2022-11-28. Retrieved 2022-11-28. ^ a b Grace, Katja; Salvatier, John; Dafoe, Allan; Zhang, Baobao; Evans, Owain (2018-07-31). "Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts". Journal of Artificial Intelligence Research. 62: 729–754. arXiv:1705.08807. doi:10.1613/jair.1.11222. ISSN 1076-9757. S2CID 8746462. Archived from the original on 2023-02-10. Retrieved 2022-11-28. ^ Zhang, Baobao; Anderljung, Markus; Kahn, Lauren; Dreksler, Noemi; Horowitz, Michael C.; Dafoe, Allan (2021-05-05). "Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers". Journal of Artificial Intelligence Research. 71. arXiv:2105.02117. doi:10.1613/jair.1.12895. ^ Stein-Perlman, Zach; Weinstein-Raun, Benjamin; Grace (2022-08-04). "2022 Expert Survey on Progress in AI". AI Impacts. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Michael, Julian; Holtzman, Ari; Parrish, Alicia; Mueller, Aaron; Wang, Alex; Chen, Angelica; Madaan, Divyam; Nangia, Nikita; Pang, Richard Yuanzhe; Phang, Jason; Bowman, Samuel R. (2022-08-26). "What Do NLP Researchers Believe? Results of the NLP Community Metasurvey". Association for Computational Linguistics. arXiv:2208.12852. ^ Markoff, John (2013-05-20). "In 1949, He Imagined an Age of Robots". The New York Times. ISSN 0362-4331. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Artificial intelligence: A handbook of professionalism. University of Sussex. January 1988. ISBN 978-0-470-21103-8. ^ Association for the Advancement of Artificial Intelligence. "AAAI Presidential Panel on Long-Term AI Futures". Archived from the original on 2022-09-01. Retrieved 2022-11-23. ^ Yampolskiy, Roman V.; Spellchecker, M. S. (2016-10-25). "Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures". arXiv:1610.07997. ^ "PT-AI 2011 – Philosophy and Theory of Artificial Intelligence (PT-AI 2011)". Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Yampolskiy, Roman V. (2013), Müller, Vincent C. (ed.), "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach", Philosophy and Theory of Artificial Intelligence, Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 5, Berlin; Heidelberg, Germany: Springer Berlin Heidelberg, pp. 389–396, doi:10.1007/978-3-642-31674-6_29, ISBN 978-3-642-31673-9, archived from the original on 2023-03-15, retrieved 2022-11-23{{citation}}: CS1 maint: work parameter with ISBN (link) ^ McLean, Scott; Read, Gemma J. M.; Thompson, Jason; Baber, Chris; Stanton, Neville A.; Salmon, Paul M. (2023-07-04). "The risks associated with Artificial General Intelligence: A systematic review". Journal of Experimental & Theoretical

Artificial Intelligence. 35 (5): 649–663. Bibcode:2023JETAI..35..649M. doi:10.1080/0952813X.2021.1964003. hdl:11343/289595. ISSN 0952-813X. S2CID 238643957. ^ Wile, Rob (August 3, 2014). "Elon Musk: Artificial Intelligence Is 'Potentially More Dangerous Than Nukes'". Business Insider. Retrieved 2024-02-22. ^ Kuo, Kaiser (2015-03-31). Baidu CEO Robin Li interviews Bill Gates and Elon Musk at the Boao Forum, March 29, 2015. Event occurs at 55:49. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Cellan-Jones, Rory (2014-12-02). "Stephen Hawking warns artificial intelligence could end mankind". BBC News. Archived from the original on 2015-10-30. Retrieved 2022-11-23. ^ Future of Life Institute. "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter". Future of Life Institute. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Future of Life Institute (October 2016). "AI Research Grants Program". Future of Life Institute. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ "SafArtInt 2016". Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Bach, Deborah (2016). "UW to host first of four White House public workshops on artificial intelligence". UW News. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Amodei, Dario; Olah, Chris; Steinhardt, Jacob; Christiano, Paul; Schulman, John; Mané, Dan (2016-07-25). "Concrete Problems in AI Safety". arXiv:1606.06565. ^ a b Future of Life Institute. "AI Principles". Future of Life Institute. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Yohsua, Bengio; Daniel, Privitera; Tamay, Besiroglu; Rishi, Bommasani; Stephen, Casper; Yejin, Choi; Danielle, Goldfarb; Hoda, Heidari; Leila, Khalatbari (May 2024). International Scientific Report on the Safety of Advanced AI (Report). Department for Science, Innovation and Technology. ^ a b Research, DeepMind Safety (2018-09-27). "Building safe artificial intelligence: specification, robustness, and assurance". Medium. Archived from the original on 2023-02-10. Retrieved 2022-11-23. ^ "SafeML ICLR 2019 Workshop". Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Browne, Ryan (2023-06-12). "British Prime Minister Rishi Sunak pitches UK as home of A.I. safety regulation as London bids to be next Silicon Valley". CNBC. Retrieved 2023-06-25. ^ Bertuzzi, Luca (October 18, 2023). "UK's AI safety summit set to highlight risk of losing human control over 'frontier' models". Euractiv. Retrieved March 2, 2024. ^ Bengio, Yoshua; Privitera, Daniel; Bommasani, Rishi; Casper, Stephen; Goldfarb, Danielle; Mavroudis, Vasilios; Khalatbari, Leila; Mazeika, Mantas; Hoda, Heidari (2024-05-17). "International Scientific Report on the Safety of Advanced AI" (PDF). GOV.UK. Archived (PDF) from the original on 2024-06-15. Retrieved 2024-07-08. Alt URL ^ Shepardson, David (1 April 2024). "US, Britain announce partnership on AI safety, testing". Retrieved 2 April 2024. ^ "What International AI Safety report says on jobs, climate, cyberwar and more". The Guardian. 2025-01-29. ISSN 0261-3077. Retrieved 2025-03-03. ^ "Launch of the First International Report on AI Safety chaired by Yoshua Bengio". mila.quebec. January 29, 2025. Retrieved 2025-03-03. ^ Goodfellow, Ian; Papernot, Nicolas; Huang, Sandy; Duan, Rocky; Abbeel, Pieter; Clark, Jack (2017-02-24). "Attacking Machine Learning with Adversarial Examples". OpenAI. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ a b Szegedy, Christian; Zaremba, Wojciech; Sutskever, Ilya; Bruna, Joan; Erhan, Dumitru; Goodfellow, Ian; Fergus, Rob (2014-02-19). "Intriguing properties of neural networks". ICLR. arXiv:1312.6199. ^ Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy (2017-02-10). "Adversarial examples in the physical world". ICLR. arXiv:1607.02533. ^ Madry, Aleksander; Makelov, Aleksandar; Schmidt, Ludwig; Tsipras, Dimitris; Vladu, Adrian (2019-09-04). "Towards Deep Learning Models Resistant to Adversarial Attacks". ICLR. arXiv:1706.06083. ^ Kannan, Harini; Kurakin, Alexey; Goodfellow, Ian (2018-03-16). "Adversarial Logit Pairing". arXiv:1803.06373. ^ Gilmer, Justin; Adams, Ryan P.; Goodfellow, Ian; Andersen, David; Dahl, George E. (2018-07-19). "Motivating the Rules of the Game for Adversarial Example Research". arXiv:1807.06732. ^ Carlini, Nicholas; Wagner, David (2018-03-29). "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text". IEEE Security and Privacy Workshops. arXiv:1801.01944. ^ Sheatsley, Ryan; Papernot, Nicolas; Weisman, Michael; Verma, Gunjan; McDaniel, Patrick (2022-09-09). "Adversarial Examples in Constrained Domains". arXiv:2011.01183. ^ Suciu, Octavian; Coull, Scott E.; Johns, Jeffrey (2019-04-13). "Exploring Adversarial Examples in Malware Detection". IEEE Security and Privacy Workshops. arXiv:1810.08280. ^ Ouyang, Long; Wu, Jeff; Jiang, Xu; Almeida, Diogo; Wainwright, Carroll L.; Mishkin, Pamela; Zhang, Chong; Agarwal, Sandhini; Slama, Katarina; Ray, Alex; Schulman, John; Hilton, Jacob; Kelton, Fraser; Miller, Luke; Simens, Maddie (2022-03-04). "Training language models to follow instructions with human feedback". NeurIPS. arXiv:2203.02155. ^ Gao, Leo; Schulman, John; Hilton, Jacob (2022-10-19). "Scaling Laws for Reward Model Overoptimization". ICML. arXiv:2210.10760. ^ Yu, Sihyun; Ahn, Sungsoo; Song, Le; Shin, Jinwoo (2021-10-27). "RoMA: Robust Model Adaptation for Offline Model-

based Optimization". NeurIPS. arXiv:2110.14188. ^ a b Hendrycks, Dan; Mazeika, Mantas (2022-09-20). "X-Risk Analysis for AI Research". arXiv:2206.05862. ^ a b "Prompt injection attacks might 'never be properly mitigated' UK NCSC warns". TechRadar. 2025-12-09. Retrieved 2025-12-12. ^ "Why Anthropic and OpenAI are obsessed with securing LLM model weights". VentureBeat. 2023-12-15. ^ "The rise of AI fake news is creating a 'misinformation superspreader'". The Washington Post. 2023-12-17. ISSN 0190-8286. Retrieved 2025-12-12. ^ Tran, Khoa A.; Kondrashova, Olga; Bradley, Andrew; Williams, Elizabeth D.; Pearson, John V.; Waddell, Nicola (2021). "Deep learning in cancer diagnosis, prognosis and treatment selection". Genome Medicine. 13 (1): 152. doi:10.1186/s13073-021-00968-x. ISSN 1756-994X. PMC 8477474. PMID 34579788. ^ Guo, Chuan; Pleiss, Geoff; Sun, Yu; Weinberger, Kilian Q. (2017-08-06). "On calibration of modern neural networks". Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research. Vol. 70. PMLR. pp. 1321–1330. ^ Ovadia, Yaniv; Fertig, Emily; Ren, Jie; Nado, Zachary; Sculley, D.; Nowozin, Sebastian; Dillon, Joshua V.; Lakshminarayanan, Balaji; Snoek, Jasper (2019-12-17). "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift". NeurIPS. arXiv:1906.02530. ^ Bogdoll, Daniel; Breitenstein, Jasmin; Heidecker, Florian; Bieshaar, Maarten; Sick, Bernhard; Fingscheidt, Tim; Zöllner, J. Marius (2021). "Description of Corner Cases in Automated Driving: Goals and Challenges". 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1023–1028. arXiv:2109.09607. doi:10.1109/ICCVW54120.2021.00119. ISBN 978-1-6654-0191-3. S2CID 237572375. ^ Hendrycks, Dan; Mazeika, Mantas; Dietterich, Thomas (2019-01-28). "Deep Anomaly Detection with Outlier Exposure". ICLR. arXiv:1812.04606. ^ Wang, Haoqi; Li, Zhizhong; Feng, Litong; Zhang, Wayne (2022-03-21). "ViM: Out-Of-Distribution with Virtual-logit Matching". CVPR. arXiv:2203.10807. ^ Hendrycks, Dan; Gimpel, Kevin (2018-10-03). "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". ICLR. arXiv:1610.02136. ^ Urbina, Fabio; Lentzos, Filippa; Invernizzi, Cédric; Ekins, Sean (2022). "Dual use of artificial-intelligence-powered drug discovery". Nature Machine Intelligence. 4 (3): 189–191. doi:10.1038/s42256-022-00465-9. ISSN 2522-5839. PMC 9544280. PMID 36211133. ^ Center for Security and Emerging Technology; Buchanan, Ben; Lohn, Andrew; Musser, Micah; Sedova, Katerina (2021). "Truth, Lies, and Automation: How Language Models Could Change Disinformation". doi:10.51593/2021ca003. S2CID 240522878. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ "Propaganda-as-a-service may be on the horizon if large language models are abused". VentureBeat. 2021-12-14. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Center for Security and Emerging Technology; Buchanan, Ben; Bansemer, John; Cary, Dakota; Lucas, Jack; Musser, Micah (2020). "Automating Cyber Attacks: Hype and Reality". Center for Security and Emerging Technology. doi:10.51593/2020ca002. S2CID 234623943. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ "Lessons Learned on Language Model Safety and Misuse". OpenAI. 2022-03-03. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Markov, Todor; Zhang, Chong; Agarwal, Sandhini; Eloundou, Tyna; Lee, Teddy; Adler, Steven; Jiang, Angela; Weng, Lilian (2022-08-10). "New-and-Improved Content Moderation Tooling". OpenAI. Archived from the original on 2023-01-11. Retrieved 2022-11-24. ^ a b Savage, Neil (2022-03-29). "Breaking into the black box of artificial intelligence". Nature. doi:10.1038/d41586-022-00858-1. PMID 35352042. S2CID 247792459. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Center for Security and Emerging Technology; Rudner, Tim; Toner, Helen (2021). "Key Concepts in AI Safety: Interpretability in Machine Learning". CSET Issue Brief. doi:10.51593/20190042. S2CID 233775541. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ McFarland, Matt (2018-03-19). "Uber pulls self-driving cars after first fatal crash of autonomous vehicle". CNNMoney. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Felder, Ryan Marshall (July 2021). "Coming to Terms with the Black Box Problem: How to Justify AI Systems in Health Care". Hastings Center Report. 51 (4): 38–45. doi:10.1002/hast.1248. ISSN 0093-0334. PMID 33821471. ^ a b Doshi-Velez, Finale; Kortz, Mason; Budish, Ryan; Bavitz, Chris; Gershman, Sam; O'Brien, David; Scott, Kate; Schieber, Stuart; Waldo, James; Weinberger, David; Weller, Adrian; Wood, Alexandra (2019-12-20). "Accountability of AI Under the Law: The Role of Explanation". arXiv:1711.01134. ^ Fong, Ruth; Vedaldi, Andrea (2017). "Interpretable Explanations of Black Boxes by Meaningful Perturbation". 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3449–3457. arXiv:1704.03296. doi:10.1109/ICCV.2017.371. ISBN 978-1-5386-1032-9. S2CID 1633753. ^ Meng, Kevin; Bau, David; Andonian, Alex; Belinkov, Yonatan (2022). "Locating and editing factual associations in GPT". Advances in Neural Information Processing

Systems. 35. arXiv:2202.05262. ^ Bau, David; Liu, Steven; Wang, Tongzhou; Zhu, Jun-Yan; Torralba, Antonio (2020-07-30). "Rewriting a Deep Generative Model". ECCV. arXiv:2007.15646. ^ Räuker, Tilman; Ho, Anson; Casper, Stephen; Hadfield-Menell, Dylan (2022-09-05). "Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks". IEEE SaTML. arXiv:2207.13243. ^ Bau, David; Zhou, Bolei; Khosla, Aditya; Oliva, Aude; Torralba, Antonio (2017-04-19). "Network Dissection: Quantifying Interpretability of Deep Visual Representations". CVPR. arXiv:1704.05796. ^ McGrath, Thomas; Kapishnikov, Andrei; Tomašev, Nenad; Pearce, Adam; Wattenberg, Martin; Hassabis, Demis; Kim, Been; Paquet, Ulrich; Kramnik, Vladimir (2022-11-22). "Acquisition of chess knowledge in AlphaZero". Proceedings of the National Academy of Sciences. 119 (47) e2206625119. arXiv:2111.09259. Bibcode:2022PNAS..11906625M. doi:10.1073/pnas.2206625119. ISSN 0027-8424. PMC 9704706. PMID 36375061. ^ Goh, Gabriel; Cammarata, Nick; Voss, Chelsea; Carter, Shan; Petrov, Michael; Schubert, Ludwig; Radford, Alec; Olah, Chris (2021). "Multimodal neurons in artificial neural networks". Distill. 6 (3). doi:10.23915/distill.00030. S2CID 233823418. ^ Olah, Chris; Cammarata, Nick; Schubert, Ludwig; Goh, Gabriel; Petrov, Michael; Carter, Shan (2020). "Zoom in: An introduction to circuits". Distill. 5 (3). doi:10.23915/distill.00024.001. S2CID 215930358. ^ Cammarata, Nick; Goh, Gabriel; Carter, Shan; Voss, Chelsea; Schubert, Ludwig; Olah, Chris (2021). "Curve circuits". Distill. 6 (1). doi:10.23915/distill.00024.006 (inactive 1 July 2025). Archived from the original on 5 December 2022. Retrieved 5 December 2022.{{cite journal}}: CS1 maint: DOI inactive as of July 2025 (link) ^ Olsson, Catherine; Elhage, Nelson; Nanda, Neel; Joseph, Nicholas; DasSarma, Nova; Henighan, Tom; Mann, Ben; Askell, Amanda; Bai, Yuntao; Chen, Anna; Conerly, Tom; Drain, Dawn; Ganguli, Deep; Hatfield-Dodds, Zac; Hernandez, Danny; Johnston, Scott; Jones, Andy; Kernion, Jackson; Lovitt, Liane; Ndousse, Kamal; Amodei, Dario; Brown, Tom; Clark, Jack; Kaplan, Jared; McCandlish, Sam; Olah, Chris (2022). "In-context learning and induction heads". Transformer Circuits Thread. arXiv:2209.11895. ^ Olah, Christopher. "Interpretability vs Neuroscience [rough note]". Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Gu, Tianyu; Dolan-Gavitt, Brendan; Garg, Siddharth (2019-03-11). "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain". arXiv:1708.06733. ^ Chen, Xinyun; Liu, Chang; Li, Bo; Lu, Kimberly; Song, Dawn (2017-12-14). "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning". arXiv:1712.05526. ^ Carlini, Nicholas; Terzis, Andreas (2022-03-28). "Poisoning and Backdooring Contrastive Learning". ICLR. arXiv:2106.09667. ^ "How 'sleeper agent' AI assistants can sabotage code". The Register. 16 January 2024. Archived from the original on 2024-12-24. Retrieved 2025-01-12. ^ a b c d Russell, Stuart J.; Norvig, Peter (2021). Artificial intelligence: A modern approach (4th ed.). Pearson. pp. 5, 1003. ISBN 978-0-13-461099-3. Retrieved September 12, 2022. ^ a b Ngo, Richard; Chan, Lawrence; Mindermann, Sören (2022). "The Alignment Problem from a Deep Learning Perspective". International Conference on Learning Representations. arXiv:2209.00626. ^ a b Pan, Alexander; Bhatia, Kush; Steinhardt, Jacob (2022-02-14). The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. International Conference on Learning Representations. Retrieved 2022-07-21. ^ Carlsmith, Joseph (2022-06-16). "Is Power-Seeking AI an Existential Risk?". arXiv:2206.13353 [cs.CY]. ^ a b c Russell, Stuart J. (2020). Human compatible: Artificial intelligence and the problem of control. Penguin Random House. ISBN 978-0-525-55863-7. OCLC 1113410915. ^ Christian, Brian (2020). The alignment problem: Machine learning and human values. W. W. Norton & Company. ISBN 978-0-393-86833-3. OCLC 1233266753. Archived from the original on February 10, 2023. Retrieved September 12, 2022. ^ Langosco, Lauro Langosco Di; Koch, Jack; Sharkey, Lee D.; Pfau, Jacob; Krueger, David (2022-06-28). "Goal Misgeneralization in Deep Reinforcement Learning". Proceedings of the 39th International Conference on Machine Learning. International Conference on Machine Learning. PMLR. pp. 12004–12019. Retrieved 2023-03-11. ^ Pillay, Tharin (2024-12-15). "New Tests Reveal AI's Capacity for Deception". TIME. Retrieved 2025-01-12. ^ Perrigo, Billy (2024-12-18). "Exclusive: New Research Shows AI Strategically Lying". TIME. Retrieved 2025-01-12. ^ a b Bommasani, Rishi; Hudson, Drew A.; Adeli, Ehsan; Altman, Russ; Arora, Simran; von Arx, Sydney; Bernstein, Michael S.; Bohg, Jeannette; Bosselut, Antoine; Brunskill, Emma; Brynjolfsson, Erik (2022-07-12). "On the Opportunities and Risks of Foundation Models". Stanford CRFM. arXiv:2108.07258. ^ Ouyang, Long; et al. (2022). "Training language models to follow instructions with human feedback" (PDF). NeurIPS. arXiv:2203.02155. ^ Zaremba, Wojciech; Brockman, Greg; OpenAI (2021-08-10). "OpenAI Codex". OpenAI. Archived from the original on February 3, 2023. Retrieved 2022-07-23. ^ Kober, Jens; Bagnell, J. Andrew; Peters, Jan (2013-09-01). "Reinforcement learning in robotics: A survey". The

International Journal of Robotics Research. 32 (11): 1238–1274. doi:10.1177/0278364913495721. ISSN 0278-3649. S2CID 1932843. Archived from the original on October 15, 2022. Retrieved September 12, 2022. ^ Knox, W. Bradley; Allievi, Alessandro; Banzhaf, Holger; Schmitt, Felix; Stone, Peter (2023-03-01). "Reward (Mis)design for autonomous driving". Artificial Intelligence. 316 103829. arXiv:2104.13906. doi:10.1016/j.artint.2022.103829. ISSN 0004-3702. S2CID 233423198. ^ Stray, Jonathan (2020). "Aligning AI Optimization to Community Well-Being". International Journal of Community Well-Being. 3 (4): 443–463. doi:10.1007/s42413-020-00086-3. ISSN 2524-5295. PMC 7610010. PMID 34723107. S2CID 226254676. ^ Russell, Stuart; Norvig, Peter (2009). Artificial Intelligence: A Modern Approach. Prentice Hall. p. 1003. ISBN 978-0-13-461099-3. ^ Smith, Craig S. "Geoff Hinton, AI's Most Famous Researcher, Warns Of 'Existential Threat'". Forbes. Retrieved 2023-05-04. ^ Bengio, Yoshua; Hinton, Geoffrey; Yao, Andrew; Song, Dawn; Abbeel, Pieter; Harari, Yuval Noah; Zhang, Ya-Qin; Xue, Lan; Shalev-Shwartz, Shai (2024). "Managing extreme AI risks amid rapid progress". Science. 384 (6698): 842–845. arXiv:2310.17688. Bibcode:2024Sci...384..842B. doi:10.1126/science.adn0117. PMID 38768279. ^ "Statement on AI Risk | CAIS". www.safe.ai. Retrieved 2024-02-11. ^ Grace, Katja; Stewart, Harlan; Sandkühler, Julia Fabienne; Thomas, Stephen; Weinstein-Raun, Ben; Brauner, Jan (2025). "Thousands of AI Authors on the Future of AI". Journal of Artificial Intelligence Research. 84. arXiv:2401.02843. doi:10.1613/jair.1.19087. ^ Perrigo, Billy (2024-02-13). "Meta's AI Chief Yann LeCun on AGI, Open-Source, and AI Risk". TIME. Retrieved 2024-06-26. ^ "What is AI alignment?". TechTarget. 2023-05-03. Retrieved 2025-06-28. ^ Ahmed, Shazeda; Ja §v‘Dska, Klaudia; Ahlawat, Archana; Winecoff, Amy; Wang, Mona (2024-04-14). "Field-building and the epistemic culture of AI safety". First Monday. doi:10.5210/fm.v29i4.13626. ISSN 1396-0466. ^ a b Ortega, Pedro A.; Maini, Vishal; DeepMind safety team (2018-09-27). "Building safe artificial intelligence: specification, robustness, and assurance". DeepMind Safety Research – Medium. Archived from the original on February 10, 2023. Retrieved 2022-07-18. ^ a b Rorvig, Mordechai (2022-04-14). "Researchers Gain New Understanding From Simple AI". Quanta Magazine. Archived from the original on February 10, 2023. Retrieved 2022-07-18. ^ Doshi-Velez, Finale; Kim, Been (2017-03-02). "Towards A Rigorous Science of Interpretable Machine Learning". arXiv:1702.08608 [stat.ML].

Wiblin, Robert (August 4, 2021). "Chris Olah on what the hell is going on inside neural networks" (Podcast). 80,000 hours. No. 107. Retrieved 2022-07-23. ^ a b Amodei, Dario; Olah, Chris; Steinhardt, Jacob; Christiano, Paul; Schulman, John; Mané, Dan (2016-06-21). "Concrete Problems in AI Safety". arXiv:1606.06565 [cs.AI]. ^ Russell, Stuart; Dewey, Daniel; Tegmark, Max (2015-12-31). "Research Priorities for Robust and Beneficial Artificial Intelligence". AI Magazine. 36 (4): 105–114. arXiv:1602.03506. doi:10.1609/aimag.v36i4.2577. hdl:1721.1/108478. ISSN 2371-9621. S2CID 8174496. Archived from the original on February 2, 2023. Retrieved September 12, 2022. ^ Wirth, Christian; Akrour, Riad; Neumann, Gerhard; Fürnkranz, Johannes (2017). "A survey of preference-based reinforcement learning methods". Journal of Machine Learning Research. 18 (136): 1–46. ^ Christiano, Paul F.; Leike, Jan; Brown, Tom B.; Martic, Miljan; Legg, Shane; Amodei, Dario (2017). "Deep reinforcement learning from human preferences". Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc. pp. 4302–4310. ISBN 978-1-5108-6096-4. ^ Heaven, Will Douglas (2022-01-27). "The new version of GPT-3 is much better behaved (and should be less toxic)". MIT Technology Review. Archived from the original on February 10, 2023. Retrieved 2022-07-18. ^ Mohseni, Sina; Wang, Haotao; Yu, Zhiding; Xiao, Chaowei; Wang, Zhangyang; Yadawa, Jay (2022-03-07). "Taxonomy of Machine Learning Safety: A Survey and Primer". ACM Computing Surveys. 55 (8): 1–38. doi:10.1145/3551385. ^ Clifton, Jesse (2020). "Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda". Center on Long-Term Risk. Archived from the original on January 1, 2023. Retrieved 2022-07-18.

Dafoe, Allan; Bachrach, Yoram; Hadfield, Gillian; Horvitz, Eric; Larson, Kate; Graepel, Thore (2021-05-06). "Cooperative AI: machines must learn to find common ground". Nature. 593 (7857): 33–36. Bibcode:2021Natur.593...33D. doi:10.1038/d41586-021-01170-0. ISSN 0028-0836. PMID 33947992. S2CID 233740521. Archived from the original on December 18, 2022. Retrieved September 12, 2022. ^ Prunkl, Carina; Whittlestone, Jess (2020-02-07). "Beyond Near- and Long-Term". Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York NY USA: ACM. pp. 138–143. doi:10.1145/3375627.3375803. ISBN 978-1-4503-7110-0. S2CID 210164673. Archived from the original on

October 16, 2022. Retrieved September 12, 2022. ^ Irving, Geoffrey; Askell, Amanda (2019-02-19). "AI Safety Needs Social Scientists". Distill. 4 (2) 10.23915/distill.00014. doi:10.23915/distill.00014. ISSN 2476-0757. S2CID 159180422. Archived from the original on February 10, 2023. Retrieved September 12, 2022. ^ Gazos, Alexandros; Kahn, James; Kusche, Isabel; Büscher, Christian; Götz, Markus (2025-04-01). "Organising AI for safety: Identifying structural vulnerabilities to guide the design of AI-enhanced socio-technical systems". Safety Science. 184 106731. doi:10.1016/j.ssci.2024.106731. ISSN 0925-7535. ^ a b c d Zwetsloot, Remco; Dafoe, Allan (2019-02-11). "Thinking About Risks From AI: Accidents, Misuse and Structure". Lawfare. Archived from the original on 2023-08-19. Retrieved 2022-11-24. ^ Zhang, Yingyu; Dong, Chuntong; Guo, Weiqun; Dai, Jiabao; Zhao, Ziming (2022). "Systems theoretic accident model and process (STAMP): A literature review". Safety Science. 152 105596. doi:10.1016/j.ssci.2021.105596. S2CID 244550153. Archived from the original on 2023-03-15. Retrieved 2022-11-28. ^ a b Gazos, Alexandros; Kahn, James; Kusche, Isabel; Büscher, Christian; Götz, Markus (2025-04-01). "Organising AI for safety: Identifying structural vulnerabilities to guide the design of AI-enhanced socio-technical systems". Safety Science. 184 106731. doi:10.1016/j.ssci.2024.106731. ISSN 0925-7535. ^ Center for Security and Emerging Technology; Hoffman, Wyatt (2021). "AI and the Future of Cyber Competition". CSET Issue Brief. doi:10.51593/2020ca007. S2CID 234245812. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ Gafni, Ruti; Levy, Yair (2024-01-01). "The role of artificial intelligence (AI) in improving technical and managerial cybersecurity tasks' efficiency". Information & Computer Security. 32 (5): 711–728. doi:10.1108/ICS-04-2024-0102. ISSN 2056-4961. ^ Abroshan, Hossein (2025). "AI to protect AI: A modular pipeline for detecting label-flipping poisoning attacks". Results in Engineering. Elsevier. doi:10.1016/j.rineng.2025.101513. ^ Abroshan, Hossein; Hashmi, Syed Waquas (2026). "A Multi-Stage Backdoor Detection (MSBD) Framework". IEEE Access. IEEE: 1–1. doi:10.1109/ACCESS.2026.3659007. ISSN 2169-3536. ^ Center for Security and Emerging Technology; Imbrie, Andrew; Kania, Elsa (2019). "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement". doi:10.51593/20190051. S2CID 240957952. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ a b Future of Life Institute (2019-03-27). AI Strategy, Policy, and Governance (Allan Dafoe). Event occurs at 22:05. Archived from the original on 2022-11-23. Retrieved 2022-11-23. ^ Zou, Andy; Xiao, Tristan; Jia, Ryan; Kwon, Joe; Mazeika, Mantas; Li, Richard; Song, Dawn; Steinhardt, Jacob; Evans, Owain; Hendrycks, Dan (2022-10-09). "Forecasting Future World Events with Neural Networks". NeurIPS. arXiv:2206.15474. ^ Gathani, Sneha; Hulsebos, Madelon; Gale, James; Haas, Peter J.; Demiralp, Ça ö F • (2022-02-08). "Augmenting Decision Making via Interactive What-If Analysis". Conference on Innovative Data Systems Research. arXiv:2109.06160. ^ Lindelauf, Roy (2021), Osinga, Frans; Sweijs, Tim (eds.), "Nuclear Deterrence in the Algorithmic Age: Game Theory Revisited", NL ARMS Netherlands Annual Review of Military Studies 2020, Nl Arms, The Hague: T.M.C. Asser Press, pp. 421–436, doi:10.1007/978-94-6265-419-8_22, ISBN 978-94-6265-418-1, S2CID 229449677{{citation}}: CS1 maint: work parameter with ISBN (link) ^ a b Newkirk II, Vann R. (2016-04-21). "Is Climate Change a Prisoner's Dilemma or a Stag Hunt?". The Atlantic. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ a b Armstrong, Stuart; Bostrom, Nick; Shulman, Carl. Racing to the Precipice: a Model of Artificial Intelligence Development (Report). Future of Humanity Institute, Oxford University. ^ a b Dafoe, Allan. AI Governance: A Research Agenda (Report). Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. ^ Dafoe, Allan; Hughes, Edward; Bachrach, Yoram; Collins, Tantum; McKee, Kevin R.; Leibo, Joel Z.; Larson, Kate; Graepel, Thore (2020-12-15). "Open Problems in Cooperative AI". NeurIPS. arXiv:2012.08630. ^ a b Dafoe, Allan; Bachrach, Yoram; Hadfield, Gillian; Horvitz, Eric; Larson, Kate; Graepel, Thore (2021). "Cooperative AI: machines must learn to find common ground". Nature. 593 (7857): 33–36. Bibcode:2021Natur.593...33D. doi:10.1038/d41586-021-01170-0. PMID 33947992. S2CID 233740521. Archived from the original on 2022-11-22. Retrieved 2022-11-24. ^ Satariano, Adam; Specia, Megan (2023-11-01). "Global Leaders Warn A.I. Could Cause 'Catastrophic' Harm". The New York Times. ISSN 0362-4331. Retrieved 2024-04-20. ^ Turchin, Alexey; Dench, David; Green, Brian Patrick (2019). "Global Solutions vs. Local Solutions for the AI Safety Problem". Big Data and Cognitive Computing. 3 (16): 1–25. doi:10.3390/bdcc3010016. ^ Crafts, Nicholas (2021-09-23). "Artificial intelligence as a general-purpose technology: an historical perspective". Oxford Review of Economic Policy. 37 (3): 521–536. doi:10.1093/oxrep/grab012. ISSN 0266-903X. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ „IOöyŽ; žÃ[PT ; _5Z•–ï; Œô_×j+ (2020-12-01). "Labor Displacement in Artificia

A Systematic Literature Review". •úpcgqNže‡f x zv[xR . 17 (2). doi:10.6163/TJEAS.202012_17(2).0002. ISSN 181
Johnson, James (2019-04-03). "Artificial intelligence & future warfare: implications for international security". Defense & Security Analysis. 35 (2): 147–169. doi:10.1080/14751798.2019.1600800. ISSN 1475-1798. S2CID 159321626. Archived from the original on 2022-11-24. Retrieved 2022-11-28. ^ Kertysova, Katarina (2018-12-12). "Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered". Security and Human Rights. 29 (1–4): 55–81. doi:10.1163/18750230-02901005. ISSN 1874-7337. S2CID 216896677. ^ Feldstein, Steven (2019). The Global Expansion of AI Surveillance. Carnegie Endowment for International Peace. ^ Agrawal, Ajay; Gans, Joshua; Goldfarb, Avi (2019). The economics of artificial intelligence: an agenda. Chicago, Illinois. ISBN 978-0-226-61347-5. OCLC 1099435014.{{cite book}}: CS1 maint: location missing publisher (link) ^ Whittlestone, Jess; Clark, Jack (2021-08-31). "Why and How Governments Should Monitor AI Development". arXiv:2108.12427. ^ a b Shevlane, Toby (2022). "Sharing Powerful AI Models | GovAI Blog". Center for the Governance of AI. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Askell, Amanda; Brundage, Miles; Hadfield, Gillian (2019-07-10). "The Role of Cooperation in Responsible AI Development". arXiv:1907.04534. ^ Gursoy, Furkan; Kakadiaris, Ioannis A. (2022-08-31), System Cards for AI-Based Decision-Making for Public Policy, arXiv:2203.04754 ^ Cobbe, Jennifer; Lee, Michelle Seng Ah; Singh, Jatinder (2021-03-01). "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems". Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery. pp. 598–609. doi:10.1145/3442188.3445921. ISBN 978-1-4503-8309-7. ^ Raji, Inioluwa Deborah; Smart, Andrew; White, Rebecca N.; Mitchell, Margaret; Gebru, Timnit; Hutchinson, Ben; Smith-Loud, Jamila; Theron, Daniel; Barnes, Parker (2020-01-27). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. New York, NY, USA: Association for Computing Machinery. pp. 33–44. doi:10.1145/3351095.3372873. ISBN 978-1-4503-6936-7. ^ Manheim, David; Martin, Sammy; Bailey, Mark; Samin, Mikhail; Greutzmacher, Ross (2025). "The necessity of AI audit standards boards". AI & Society. 40 (8): 6609–6624. arXiv:2404.13060. doi:10.1007/s00146-025-02320-y. ^ Novelli, Claudio; Taddeo, Mariarosaria; Floridi, Luciano (2024). "Accountability in artificial intelligence: what it is and how it works". AI & Society. 39 (4): 1871–1882. doi:10.1007/s00146-023-01635-y. hdl:11585/914099. ^ Manheim, David (26 June 2023). "Building a Culture of Safety for AI: Perspectives and Challenges". SSRN 4491421. ^ "NeMo Guardrails". NVIDIA NeMo Guardrails. Retrieved 2024-12-08. ^ "Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations". Meta AI. Retrieved 2024-12-08. ^ Šekrst, Kristina; McHugh, Jeremy; Cefalu, Jonathan Rodriguez (2024). "AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development". arXiv:2411.14442 [cs.CY]. ^ Dong, Yi; Mu, Ronghui; Jin, Gaojie; Qi, Yi; Hu, Jinwei; Zhao, Xingyu; Meng, Jie; Ruan, Wenjie; Huang, Xiaowei (2024). "Building Guardrails for Large Language Models". arXiv:2402.01822 [cs]. ^ D'Alessandro, W. (2024). "Deontology and safe artificial intelligence". Philosophical Studies. 182 (7): 1681–1704. doi:10.1007/s11098-024-02174-y. ^ D'Alessandro, William; Kirk-Giannini, Chad D. (2025). "Artificial Intelligence: Approaches to Safety". Philosophy Compass. 20 (5) e70039. doi:10.1111/phc3.70039. ^ Ziegler, Bart (8 April 2022). "Is It Time to Regulate AI?". Wall Street Journal. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Reed, Chris (2018-09-13). "How should we regulate artificial intelligence?". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 376 (2128) 20170360. Bibcode:2018RSPTA.37670360R. doi:10.1098/rsta.2017.0360. ISSN 1364-503X. PMC 6107539. PMID 30082306. ^ Belton, Keith B. (2019-03-07). "How Should AI Be Regulated?". IndustryWeek. Archived from the original on 2022-01-29. Retrieved 2022-11-24. ^ National Security Commission on Artificial Intelligence (2021), Final Report ^ National Institute of Standards and Technology (2021-07-12). "AI Risk Management Framework". NIST. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Richardson, Tim (2021). "Britain publishes 10-year National Artificial Intelligence Strategy". Archived from the original on 2023-02-10. Retrieved 2022-11-24. ^ a b "Guidance: National AI Strategy". GOV.UK. 2021. Archived from the original on 2023-02-10. Retrieved 2022-11-24. ^ Hardcastle, Kimberley (2023-08-23). "We're talking about AI a lot right now – and it's not a moment too soon". The Conversation. Retrieved 2023-10-31. ^ "Iconic Bletchley Park to host UK AI Safety Summit in early November". GOV.UK. Retrieved 2023-10-31. ^ Colville, Alex (2025-07-30). "How China Sees AI Safety".

China Media Project. Retrieved 2025-08-09. ^ Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity. "IARPA – TrojAI". Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Turek, Matt. "Explainable Artificial Intelligence". Archived from the original on 2021-02-19. Retrieved 2022-11-24. ^ Draper, Bruce. "Guaranteeing AI Robustness Against Deception". Defense Advanced Research Projects Agency. Archived from the original on 2023-01-09. Retrieved 2022-11-24. ^ National Science Foundation (23 February 2023). "Safe Learning-Enabled Systems". Archived from the original on 2023-02-26. Retrieved 2023-02-27. ^ "General Assembly adopts landmark resolution on artificial intelligence". UN News. 21 March 2024. Archived from the original on 20 April 2024. Retrieved 21 April 2024. ^ Say, Mark (23 May 2024). "DSIT announces funding for research on AI safety". Archived from the original on 24 May 2024. Retrieved 11 June 2024. ^ Mäntymäki, Matti; Minkkinen, Matti; Birkstedt, Teemu; Viljanen, Mika (2022). "Defining organizational AI governance". AI and Ethics. 2 (4): 603–609. doi:10.1007/s43681-022-00143-x. ISSN 2730-5953. S2CID 247119668. ^ a b c Brundage, Miles; Avin, Shahar; Wang, Jasmine; Belfield, Haydn; Krueger, Gretchen; Hadfield, Gillian; Khlaaf, Heidy; Yang, Jingying; Toner, Helen; Fong, Ruth; Maharaj, Tegan; Koh, Pang Wei; Hooker, Sara; Leung, Jade; Trask, Andrew (2020-04-20). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims". arXiv:2004.07213. ^ "Welcome to the Artificial Intelligence Incident Database". Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ Wiblin, Robert; Harris, Keiran (2022). "Nova DasSarma on why information security may be critical to the safe development of AI systems". 80,000 Hours. Archived from the original on 2022-11-24. Retrieved 2022-11-24. ^ OpenAI (2022-06-02). "Best Practices for Deploying Language Models". OpenAI. Archived from the original on 2023-03-15. Retrieved 2022-11-24. ^ OpenAI. "OpenAI Charter". OpenAI. Archived from the original on 2021-03-04. Retrieved 2022-11-24. ^ Future of Life Institute (2016). "Autonomous Weapons Open Letter: AI & Robotics Researchers". Future of Life Institute. Retrieved 2022-11-24. Unsolved Problems in ML Safety

On the Opportunities and Risks of Foundation Models

An Overview of Catastrophic AI Risks

AI Accidents: An Emerging Threat

Engineering a Safer World

%¼ END OF SOURCE 1

AI Safety Institute

Type: Ø<ß Web Page
URL: https://en.wikipedia.org/wiki/AI_Safety_Institute
Date Extracted: 2026-02-05T17:27:39.582Z
Word Count: 1,835 words

From Wikipedia, the free encyclopedia An AI Safety Institute (AISI) is a state-backed institute aiming to evaluate and ensure the safety of advanced artificial intelligence (AI) models, also called frontier AI models.[1]

AI safety gained prominence in 2023, notably with public declarations about potential existential risks from AI. During the AI Safety Summit in November 2023, the United Kingdom and the United States both created their own AISI. During the AI Seoul Summit in May 2024, international leaders agreed to form a network of AI Safety Institutes, comprising institutes from the UK, the US, Japan, France, Germany, Italy, Singapore, South Korea, Australia, Canada and the European Union.[2] In 2025, the UK's AI Safety Institute was renamed the "AI Security Institute", and its US counterpart became the Center for AI Standards and Innovation (CAISI). In 2023, Rishi Sunak, the Prime Minister of the United Kingdom, expressed his intention to "make the UK not just the intellectual home but the geographical home of global AI safety regulation" and unveiled plans for an AI Safety Summit.[3] He emphasized the need for independent safety evaluations, stating that AI companies cannot "mark their own homework".[4] During the summit in November 2023, the UK AISI was officially established as an evolution of the Frontier AI Taskforce,[5] and the US AISI as part of the NIST. Japan followed by launching an AI safety institute in February 2024.[6]

Politico reported in April 2024 that many AI companies had not shared pre-deployment access to their most advanced AI models for evaluation. Meta's president of global affairs Nick Clegg said that many AI companies were waiting for the UK and the US AI Safety Institutes to work out common evaluation rules and procedures.[7] An agreement was indeed concluded between the UK and the US in April 2024 to collaborate on at least one joint safety test.[8] Initially established in London, the UK AI Safety Institute announced in May 2024 that it would open an office in San Francisco, where many AI companies are located. This is part of a plan to "set new, international standards on AI safety", according to UK's technology minister Michele Donelan.[9][10]

At the AI Seoul Summit in May 2024, the European Union and other countries agreed to create their own AI safety institutes, forming an international network.[2] The United Kingdom founded in April 2023 a safety organisation called Frontier AI Taskforce, with an initial budget of £100 million.[11] In November 2023, it evolved into the AI Safety Institute, and continued to be led by Ian Hogarth. The AISI is part of the United Kingdom's Department for Science, Innovation and Technology.[5]

The United Kingdom's AI strategy aims to balance safety and innovation. Unlike the European Union which adopted the AI Act, the UK is reluctant to legislate early, considering that it may lower the sector's growth, and that laws might be rendered obsolete by technological progress.[6]

In May 2024, the institute open-sourced an AI safety tool called "Inspect", which evaluates AI model capabilities such as reasoning and their degree of autonomy.[12]

In February 2025, the UK body was renamed the AI Security Institute. Observers saw the name change as a signal that the institute will not focus on ethical issues such as algorithmic bias or freedom of speech in AI applications.[13] The US AISI was founded in November 2023 as part of the NIST. This happened the day after the signature of the Executive Order 14110.[14] In February 2024, Joe Biden's former economic policy adviser Elizabeth Kelly was appointed to lead it.[15]

In February 2024, the US government created the US AI Safety Institute Consortium (AISIC), regrouping more than 200 organizations such as Google, Anthropic or Microsoft.[16]

In March 2024, a budget of $10 million was allocated.[17] Observers noted that this investment is relatively small, especially considering the presence of many big AI companies in the US. The NIST itself, which hosts the AISI, is also known for its chronic lack of funding.[18][6] Biden administration's request for additional funding was met with further budget cuts from congressional appropriators.[19][18]

Under President Trump, plans for members of the agency to attend the February AI Action Summit in Paris February 2025 were scrapped.[20] The US and the UK refused to sign the summit's final communique. US Vice President JD Vance said "pro-growth AI policies" should be prioritised over safety.[21]

The name of the agency was changed in June 2025 to the Center for AI Standards and Innovation (CAISI) and its mission transformed.[22] According to Secretary of Commerce Howard Lutnick, "For far too long, censorship and regulations have been used under the guise of national security. Innovators will no longer be limited by these standards. CAISI will evaluate and enhance US innovation of these rapidly developing commercial AI systems while ensuring they remain secure to our national security standards."[23][24] The US Department of Commerce stated that CAISI would represent American interests internationally, guarding against burdensome and unnecessary regulation of US technologies by foreign governments. It collaborates with the NIST Information Technology Laboratory.[24] The Ministry of Electronics and Information Technology held consultations with Meta Platforms, Google, Microsoft, IBM, OpenAI, NASSCOM, Broadband India Forum, Software Alliance, Indian Institutes of Technology (IITs), The Quantum Hub, Digital Empowerment Foundation, and Access Now on October 7, 2024, in relation to the establishment of the AI Safety Institute. The decision was made to shift focus from regulation to standards-setting, risk identification, and damage detection—all of which require interoperable technologies. The AISI may spend the ¹20 crore allotted to the Safe and Trusted Pillar of the IndiaAI Mission for the initial budget. Future funding may come from other components of the IndiaAI Mission.[25][26]

UNESCO and MeitY began consulting on AI Readiness Assessment Methodology under Safety and Ethics in Artificial Intelligence from 2024. It is to encourage the ethical and responsible use of AI in industries. The study will find areas where government can become involved, especially in attempts to strengthen institutional and regulatory capabilities.[27][28]

Minister for Electronics & Information Technology Ashwini Vaishnaw announced the creation of an IndiaAI Safety Institute on January 30, 2025, to ensure the ethical and safe application of AI models. The institute will promote domestic R&D that is grounded in India's social, economic, cultural, and linguistic diversity and is based on Indian datasets. With the help of academic and research institutions, as well as private sector partners, the institute will follow the hub-and-spoke approach to carry out projects within Safe and Trusted Pillar of the IndiaAI Mission.[29][30] It operates under a "hub-and-spoke" model with collaboration from academic institutions (e.g., IITs), tech firms, and international organizations like UNESCO.[31] Alignment Research Center

Foundation model

Regulation of artificial intelligence ^ "Safety institutes to form 'international network' to boost AI research and tests". The Independent. 2024-05-21. Retrieved 2024-07-06. ^ a b Desmarais, Anna (2024-05-22). "World leaders agree to launch network of AI safety institutes". euronews. Retrieved 2024-06-15. ^ Browne, Ryan (2023-06-12). "British Prime Minister Rishi Sunak pitches UK as home of A.I. safety regulation as London bids to be next Silicon Valley". CNBC. Retrieved 2024-06-21. ^ "Rishi Sunak: AI firms cannot 'mark their own homework'". BBC. 2023-11-01. Retrieved 2024-06-21. ^ a b "Introducing the AI Safety Institute". GOV.UK. November 2023. Retrieved 2024-06-15. ^ a b c Henshall, Will (April 1, 2024). "U.S., U.K. Announce Partnership to Safety Test AI Models". TIME. Retrieved 2024-07-06. ^ "Rishi Sunak promised to make AI safe. Big Tech's not playing ball". Politico. 2024-04-26. Retrieved 2024-06-15. ^ David, Emilia (2024-04-02). "US and UK will work together to test AI models for safety threats". The Verge. Retrieved 2024-06-21. ^ Coulter, Martin (20 May 2024). "Britain's AI safety institute to open US office". Reuters. ^ Browne, Ryan (2024-05-20).

"Britain expands AI Safety Institute to San Francisco amid scrutiny over regulatory shortcomings". CNBC. Retrieved 2024-06-15. ^ "Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI". GOV.UK. Retrieved 2024-07-06. ^ Wodecki, Ben (May 15, 2024). "AI Safety Institute Launches AI Model Safety Testing Tool Platform". AI Business. ^ Poireault, Kevin (2025-02-14). "UK's AI Safety Institute Rebrands Amid Government Strategy Shift". Infosecurity Magazine. Retrieved 2025-07-23. ^ Henshall, Will (2023-11-01). "Why Biden's AI Executive Order Only Goes So Far". TIME. Retrieved 2024-07-07. ^ Henshall, Will (2024-02-07). "Biden Economic Adviser Elizabeth Kelly Picked to Lead AI Safety Testing Body". TIME. Retrieved 2024-07-06. ^ Shepardson, David (February 8, 2024). "US says leading AI companies join safety consortium to address risks". Reuters. ^ "Majority Leader Schumer Announces First-Of-Its-Kind Funding To Establish A U.S. Artificial Intelligence Safety Institute; Funding Is A Down Payment On Balancing Safety With AI Innovation And Will Aid Development Standards, Tools, And Tests To Ensure AI Systems Operate Safely". www.democrats.senate.gov. 2024-03-07. Retrieved 2024-07-06. ^ a b Zakrzewski, Cat (2024-03-08). "This agency is tasked with keeping AI safe. Its offices are crumbling". Washington Post. ISSN 0190-8286. Retrieved 2024-07-06. ^ "NIST would 'have to consider' workforce reductions if appropriations cut goes through". FedScoop. 2024-05-24. Retrieved 2024-07-06. ^ Dastin, Jeffrey (2025-02-07). "Exclusive: Trump's Paris AI summit delegation won't include AI Safety Institute staff, sources say". Reuters. Retrieved 2025-08-29. ^ "UK and US refuse to sign international AI declaration". BBC News. 2025-02-11. Retrieved 2025-08-29. ^ Robertson, Adi (2025-06-04). "US removes 'safety' from AI Safety Institute". The Verge. Retrieved 2025-08-25. ^ "Trump administration cuts 'Safety' from AI Safety Institute". NBC News. 2025-06-04. Retrieved 2025-08-25. ^ a b "Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation". US Department of Commerce. June 3, 2025. Retrieved August 15, 2025. This article incorporates text from this source, which is in the public domain. ^ "Govt mulls setting up Artificial Intelligence Safety Institute". Hindustan Times. 2024-10-13. Archived from the original on 2024-11-20. Retrieved 2025-02-17. ^ Jeevanandam, Nivash (15 October 2024). "MeitY Hosts Consultation for Establishing India AI Safety Institute under IndiaAI Mission's Safe and Trusted Pillar". IndiaAI. Retrieved 2025-02-17. ^ "UNESCO and the Ministry of Electronics and Information Technology, Host Multi-Stakeholder Consultation on Safety and Ethics in Artificial Intelligence". Press Information Bureau. Ministry of Electronics & IT, Government of India. 16 November 2024. Retrieved 19 February 2025. ^ "UNESCO and Ministry of Electronics and Information Technology (MeitY) host stakeholder consultation on AI Readiness Assessment Methodology (RAM) in India". Press Information Bureau. Ministry of Electronics & IT, Government of India. 21 January 2025. Retrieved 19 February 2025. ^ "With robust and high end Common computing facility in place, India all set to launch its own safe & secure indigenous AI model at affordable cost soon: Shri Ashwini Vaishnaw". Press Information Bureau. Ministry of Electronics & IT, Government of India. 30 January 2025. Retrieved 24 February 2025. ^ Kumar, Animesh (5 February 2025). "India's AI Safety Institute: The Role Of AISI In The Dynamic AI Landscape". Mondaq. Retrieved 2025-02-24. ^ "India AI Safety Institute and policy initiatives". Press Information Bureau. 30 January 2025. Canada AI Safety Institute

European AI Office

Japan AI Safety Institute

Singapore AI Safety Institute

South Korea AI Safety Institute

UK AI Security Institute

US AI Safety Institute