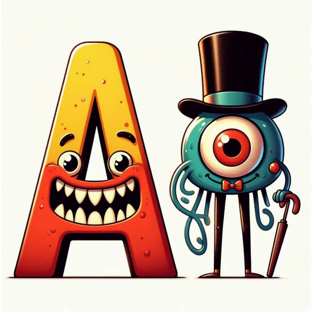
What problems does AI pose? (A.I. - What to Worry About?)	

An AI Generated Overview of Possible Challenges and Threats Posed by AI - - -
(WARNING - Hallucinations are possible)

ChatGPT 4o - 11/24/2024 … Possible Challenges and Threats Posed by AI.
(Prompt - Please list twelve challenges or threats posed by artificial intelligence in a manner that would be appropriate for first-year college students.Please make your response roughly 1000-1200 words in length and appropriate for a college audience.)

Twelve Challenges and Threats Posed by Artificial Intelligence

Artificial intelligence (AI) is transforming the world, from reshaping industries to enhancing our daily lives. Yet, with its rapid development and widespread application come significant challenges and threats. While many of these issues are technical or economic, others affect society, ethics, and even our sense of humanity. Below, we explore twelve key challenges posed by AI, each critical for understanding the full implications of this technology.

---

1. Bias and Fairness

AI systems often reflect the biases present in their training data, which can lead to unfair or discriminatory outcomes. For example, facial recognition systems have been shown to perform poorly on individuals with darker skin tones, often misidentifying them at a much higher rate than lighter-skinned individuals. This bias isn't just a technical flaw—it can have serious real-world consequences, such as unfair treatment in hiring processes or law enforcement. Ensuring fairness requires careful design, diverse datasets, and ongoing scrutiny of AI applications.

2. Privacy Violations

AI systems often rely on massive amounts of personal data to function effectively. From online shopping recommendations to facial recognition in public spaces, the collection and analysis of personal data pose significant privacy concerns. For instance, AI algorithms can predict sensitive information about individuals, such as their health status or political affiliations, based on seemingly unrelated data. The use of such data without consent infringes on privacy and raises questions about how much surveillance is acceptable in society.

3. Job Displacement

Automation driven by AI has already begun to replace jobs in manufacturing, transportation, and even customer service. Self-driving vehicles, for example, threaten the livelihoods of millions of truck drivers, while automated checkout systems reduce the need for cashiers. While AI can create new job opportunities, the transition often leaves displaced workers struggling to adapt. Addressing this challenge requires significant investment in education, retraining programs, and social safety nets.

4. Weaponization of AI

AI's potential use in military and defense applications is both promising and alarming. Autonomous drones and AI-powered weapon systems could execute decisions faster than human operators, but their lack of moral judgment introduces ethical dilemmas. Moreover, such technologies could be used irresponsibly or fall into the hands of malicious actors, escalating conflicts or enabling acts of terrorism.

5. Lack of Transparency

Many AI systems function as "black boxes," meaning their decision-making processes are not easily understood by humans, including the engineers who build them. This lack of transparency can make it difficult to detect errors, biases, or harmful behaviors. For example, if an AI denies someone a loan or parole, it may not be clear why the decision was made. Increasing transparency through explainable AI is an ongoing challenge for researchers and developers.

6. Misinformation and Deepfakes

AI makes it easier than ever to create and distribute fake content, including convincing videos, audio, and images. Deepfakes, which use AI to manipulate media, can make people appear to say or do things they never did. Such technologies undermine trust in legitimate information sources and can be used for malicious purposes, such as political manipulation or financial scams.

7. Ethical Dilemmas

AI systems often face decisions with no clear "right" answer. For instance, an autonomous vehicle might have to choose between hitting a pedestrian or swerving into oncoming traffic. Such dilemmas force developers to encode ethical principles into AI systems, but deciding whose values to prioritize is highly contentious. These ethical challenges become even more complex when AI systems operate in culturally diverse environments.

8. Monopolization of Power

The development and deployment of advanced AI require substantial resources, including expertise, data, and computing power. As a result, a small number of tech companies and governments dominate the AI landscape. This concentration of power raises concerns about monopolistic behavior, lack of competition, and potential misuse of AI for political or economic control.

9. Erosion of Human Autonomy

AI systems increasingly influence the choices we make, from the products we buy to the news we consume. Algorithms that optimize for engagement can push users toward content that reinforces their existing beliefs, creating echo chambers and polarizing society. Over time, this erosion of human autonomy may lead individuals to rely on AI for decision-making, undermining critical thinking and personal agency.

10. Safety and Reliability

Ensuring that AI systems behave safely and reliably, especially in high-stakes environments, is a major technical challenge. For example, an autonomous vehicle must be able to navigate complex traffic scenarios without causing accidents. Similarly, AI systems used in healthcare must provide accurate diagnoses to avoid putting patients at risk. Failures in safety and reliability can have catastrophic consequences, highlighting the need for rigorous testing and oversight.

11. Economic Inequality

While AI has the potential to boost productivity and economic growth, its benefits are not evenly distributed. Wealthy individuals and organizations with access to advanced AI can leverage it to gain even more advantages, widening the gap between rich and poor. Furthermore, regions without the infrastructure or expertise to develop AI may fall behind in the global economy, exacerbating inequality on an international scale.

12. Existential Risks

Finally, some researchers and thinkers, including prominent figures like Elon Musk and Nick Bostrom, have warned about the long-term risks posed by superintelligent AI. If AI systems were to surpass human intelligence and act in ways misaligned with our values, they could pose existential threats to humanity. While such scenarios may seem far-fetched, they underscore the importance of careful oversight and international cooperation in AI research.

Addressing the Challenges

Understanding these challenges is the first step toward addressing them. Governments, organizations, and individuals all have roles to play in ensuring AI is developed responsibly:

1. **Policy and Regulation:** Governments must create laws and regulations to address issues like privacy, fairness, and safety.
2. **Education and Awareness:** Equipping individuals with the knowledge to understand and interact with AI responsibly is crucial.
3. **International Collaboration:** AI is a global technology, and solving its challenges requires cooperation across borders.
4. **Ethical Design:** Developers must prioritize ethical considerations when designing and deploying AI systems.

Conclusion

AI holds immense potential to improve lives and solve some of humanity's biggest problems. However, its rapid development also introduces significant challenges and threats. Addressing these issues requires a multidisciplinary approach that combines technical expertise with ethical reflection, societal input, and robust governance. As first-year college students, you are in a unique position to shape the future of AI—whether through innovation, advocacy, or informed dialogue. By understanding these challenges, you can contribute to ensuring AI serves as a force for good in the world.

Human Views on the Possible Challenges and Threats Posed by AI - - -
(WARNING - Humans are fallible)

Shorter Podcasts and Videos:

60% of USA Jobs Could Be Impacted by AI
AI Daily Brief - Nathaniel Whittemore (8:12)
(January 16, 2024) [Video]
https://www.youtube.com/watch?v=YqzK034CMwo

The A.I. Dilemma:
    1st Contact: Social Media (Maximize Engagement)
    2nd Contact: Generative AI (Creation AI)
Your Undivided Attention - Tristin Harris, Aza Raskin
Center for Humane Technology (5:45) (**7:01 to 12:46**)
(April 5, 2023) [Video] https://www.youtube.com/watch?v=xoVJKj8lcNQ&t=421

The AI Doomsday Scenario: A Comprehensive Guide to P(doom)
A Beginner's Guide to AI (22:09)
(June 26, 2024) [Podcast]
https://www.listennotes.com/podcasts/a-beginners-guide/the-ai-doomsday-scenario-a-QXa4MnI8HdK/

AI is Killing the Open Web
The AI Daily Brief - Nathaniel Whittemore (8:58)
(2023, July 6) [Video/Podcast]
https://www.youtube.com/watch?v=ESzW-mOS3eE

AI will take your job.
UBI Works - Mark Cuban and Bloomberg (4:15)
(March, 2024) [Video]
https://www.youtube.com/watch?v=UXgJyDNBPrY

The dark side of competition in AI
TED - Liv Boree - (12:02)
(October, 2023) [Video]
https://www.ted.com/talks/liv_boeree_the_dark_side_of_competition_in_ai

AI Researchers Put Risk of Human Extinction from AI at 5%
The AI Daily Brief - Nathaniel Whittemore (9:14)
(January 4, 2024) [Video]
https://www.youtube.com/watch?v=2FtRxGYeRio&t=197s

Are AI Girlfriends A Threat to Society?
The AI Daily Brief - Nathaniel Whittemore (8:59)
(December 18, 2023) [Video]
https://www.youtube.com/watch?v=UFZMUsTOXvE

Doomsday AI: When Silicon Overlords Decide Our Fate
A Beginner's Guide to AI (17:48)
(March 1, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/doomsday-ai-when-silicon-overlords-decide-our-fate/id1701165010?i=1000647714603

Doomsday AI: When Silicon Overlords Decide Our Fate
A Beginner's Guide to AI (13:12)
(April 12, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/from-paperclips-to-disaster-ais-unseen-risks/id1701165010?i=1000652250634

Is AI Dangerous? Another AI Pioneer Sounds the Warning
The AI Daily Brief - Nathaniel Whittemore (5:10)
(May 18, 2023) [Video]
https://www.youtube.com/watch?v=8_fj13WHFgY

[UNI only] **Technology, Power, and the Twenty-First Century's Greatest Dilemma** Audio clip from book "The Coming Wave" by Mustafa Suleyman with Michael Bhaskar - read by Mustafa Suleyman. Found at the end of book review Metz, Cade. (2023, December 29). "Audiobook of the week: How to 'contain' the threat of A.I.? New York Times. (5:34).
https://www.nytimes.com/2023/12/29/books/review/mustafa-suleyman-the-coming-wave.html

Why Are People So Scared of Sora?
The AI Daily Brief - Nathaniel Whittemore (13:12)
(February 20, 2024) [Video]
https://www.youtube.com/watch?v=Rn3XkhaPD5A

Why the IMF Thinks AI Will Increase Inequality
The AI Daily Brief - Nathaniel Whittemore (4:54)
(June 19, 2024) [Video]
https://www.youtube.com/watch?v=L3lzvnGzE_U

Will superintelligent AI end the world?
TED - Eliezer Yudkowsky - (10:32) (first five minutes)
(April, 2023) [Video]
https://www.ted.com/talks/eliezer_yudkowsky_will_superintelligent_ai_end_the_world

Longer podcasts and videos:

AI Scouting Report: NOT Investment Advice Edition.
Cognitive Revolution - Nathan Labenz (1:01:55)
(June 1, 2024) [Podcast]
https://www.cognitiverevolution.ai/ai-scouting-report-not-investment-advice-edition/

Catch Up on the State of the AI Industry.
Decoder - Nilay Patel, Alex Heath, & Kylie Robinson (45:31)
(June 21, 2024) [Podcast]
https://www.theverge.com/2024/6/21/24183510/catch-up-on-the-state-of-the-ai-industry

Clearview AI and the end of privacy, with author Kashmir Hill
Decoder - Nilay Patel and Kashmir Hill (1:01:49)
(October 17, 2023) [Podcast]
https://www.theverge.com/23919134/kashmir-hill-your-face-belongs-to-us-clearview-ai-facial-recognition-privacy-decoder

Connor Leahy: The Unspoken Risks of Centralizing AI Power.
Eye on AI - Craig Smith and Connor Leahy - (1:00:02)
(2023, November 29) [Video]
https://www.youtube.com/watch?v=BhQBmVZ5XP4

Deepfake Abuse is a Crisis w/ Kat Tenbarge
Tech Won't Save Us - Paris Marx and Kat Tenbarge (1:08:43)
(April 4, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/deepfake-abuse-is-a-crisis-w-kat-tenbarge/id1507621076?i=1000651396180

Eliezer Yudkowsky - Why AI Will Kill Us, Aligning LLMs, Nature of Intelligence, SciFi, & Rationality.
Dwarkesh Podcast - Dwarkesh Patel and Eliezer Yudkowsky (4:03:25)
(April 6, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/eliezer-yudkowsky-why-ai-will-kill-us-aligning-llms/id1516093381?i=1000607719339

Generative AI is a Climate Disaster w/ Sasha Luccioni
Tech Won't Save Us - Paris Marx and Sasha Luccioni (1:00:00)
(July 18, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/tech-wont-save-us/id1507621076

How AI Disrupts Elections and Influences Voter Choices
Tech Tank - Darrell West and Sarah M. L. Bender (33:42)
(June 5, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/how-ai-disrupts-elections-and-influences-voter-choices/id1526725061?i=1000615683223

Is AI just all hype?
The TED AI Show - Bilawal Sidhu and Gary Marcus (39:04)
(July 8, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/is-ai-just-all-hype-w-gary-marcus/id1741574582

The Man Who Predicted the Downfall of Thinking
Your Undivided Attention - Tristin Harris, Sean Illing, and Lance Strate
Center for Humane Technology (59:33)
(March 6, 2022) [Video]
https://www.youtube.com/watch?v=WPr9h-yb1rU&list=PL3Pb2mJQGJV7TWAs2A4vl4hjgtWCLsDwd&index=1

The Most Amazing — and Dangerous — Technology in the World
The Ezra Klein Show - New York Times - Ezra Klein & Chris Miller (58:28)
(April 4, 2023) [Podcast].
https://podcasts.apple.com/us/podcast/the-most-amazing-and-dangerous-technology-in-the-world/id1548604447?i=1000607306353

No one is immune to AI harms with Dr. Joy Buolamwini.
Your Undivided Attention - Tristan Harris, Asa Raskin, & Joy Buolamwini (47:46)
(October 26, 2023) [Podcast]
https://www.humanetech.com/podcast/no-one-is-immune-to-ai-harms-with-dr-joy-buolamwini

Protecting Our Freedom of Thought with Nita Farahany
Your Undivided Attention - The Center for Humane Technology
Tristan Harris, Aza Raskin, and Nita Farahany (44:07)
(August 3, 2023) [Podcast]
https://www.humanetech.com/podcast/protecting-our-freedom-of-thought-with-nita-farahany

This Moment in AI: How We Got Here and Where We're Going.
Your Undivided Attention - Sasha Fagan, Tristin Harris, Aza Raskin
Center for Humane Technology (36:55)
(August 12, 2024) [Podcast]
https://www.humanetech.com/podcast/this-moment-in-ai-how-we-got-here-and-where-were-going

Tom Davidson on How Quickly AI Could Automate the Economy
Future of Life - Gus Docker & Tom Davidson (1:56:22)
(September 8, 2023 [Podcast]
https://futureoflife.org/podcast/tom-davidson-on-how-quickly-ai-could-automate-the-economy/

Tom Davidson on How Quickly AI Could Transform the World
80,000 Hours - Luisa Rodriguez & Tom Davidson (3:01:58) yikes!
(May 5, 2023) [Podcast]
https://80000hours.org/podcast/episodes/tom-davidson-how-quickly-ai-could-transform-the-world/

Why AI is a Threat to Artists w/ Molly Crabapple
Tech Won't Save Us - Paris Marx and Molly Crabapple (54:47)
(June 29, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/why-ai-is-a-threat-to-artists-w-molly-crabapple/id1507621076?i=1000618717251

Magazine, Journal and Newspaper Articles:

[UNI only] Angwin, Julia. (2024, May 15). **Press pause on the Silicon Valley hype machine**. *New York Times*. https://www.nytimes.com/2024/05/15/opinion/artificial-intelligence-ai-openai-chatgpt-overrated-hype.html?searchResultPosition=1

[UNI only] Belle, Lin. (2024, February). **Bad chatbots pose new threat on web**. *Wall Street Journal*. https://login.proxy.lib.uni.edu/login?url=https://www.proquest.com/newspapers/bad-chatbots-pose-new-threat-on-web/docview/2932785049/se-2?accountid=14691

[UNI only] Hill, Kashmir, & Hsu, Tiffany. (2024, June 10). **It looked like a reliable news site. It was an A.I. chop shop**. *New York Times*. https://www.nytimes.com/2024/06/06/technology/bnn-breaking-ai-generated-news.html

[UNI only] Hoffman, Benjamin. (2024, June 11). **First Came 'Spam.' Now, With A.I., We've Got 'Slop'.** *New York Times*. https://www.nytimes.com/2024/06/11/style/ai-search-slop.html

Jo, A. (2023). **The promise and peril of generative AI. Researchers are excited but apprehensive about the latest advances in artificial intelligence.** *Nature, 614*(1), 214-216. https://www.nature.com/articles/d41586-023-00340-6

[UNI only] Metz, Cade. (2023, May 1). **What exactly are the dangers posed by A.I.?** *New York Times*. https://www.nytimes.com/2023/05/01/technology/ai-problems-danger-chatgpt.html

Scholarly Journal, Preprint and Conference Articles

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., ... & Mindermann, S. (2024). **Managing extreme AI risks amid rapid progress**. *Science, 384*(6698), 842-845. https://www.science.org/doi/full/10.1126/science.adn0117

Abstract: Artificial intelligence (AI) is progressing rapidly, and companies are shifting their focus to developing generalist AI systems that can autonomously act and pursue goals. Increases in capabilities and autonomy may soon massively amplify AI's impact, with risks that include large-scale social harms, malicious uses, and an irreversible loss of human control over autonomous AI systems. Although researchers have warned of extreme risks from AI (1), there is a lack of consensus about how to manage them. Society's response, despite promising first steps, is incommensurate with the possibility of rapid, transformative progress that is expected by many experts. AI safety research is lagging. Present governance initiatives lack the mechanisms and institutions to prevent misuse and recklessness and barely address autonomous systems. Drawing on lessons learned from other safety-critical technologies, we outline a comprehensive plan that combines technical research and development (R&D) with proactive, adaptive governance mechanisms for a more commensurate preparation.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). **An overview of catastrophic AI risks**. *arXiv* preprint arXiv:2306.12001. https://arxiv.org/pdf/2306.12001

Abstract: Rapid advancements in artificial intelligence (AI) have sparked growing concerns among experts, policymakers, and world leaders regarding the potential for increasingly advanced AI systems to pose catastrophic risks. Although numerous risks have been detailed separately, there is a pressing need for a systematic discussion and illustration of the potential dangers to better inform efforts to mitigate them. This paper provides an overview of the main sources of catastrophic AI risks, which we organize into four categories: malicious use, in which individuals or groups intentionally use AIs to cause harm; AI race, in which competitive environments compel actors to deploy unsafe AIs or cede control to AIs; organizational risks, highlighting how human factors and complex systems can increase the chances of catastrophic accidents; and rogue AIs, describing the inherent difficulty in controlling agents far more intelligent than humans. For each category of risk, we describe specific hazards, present illustrative stories, envision ideal scenarios, and propose practical suggestions for mitigating these dangers. Our goal is to foster a comprehensive understanding of these risks and inspire collective and proactive efforts to ensure that AIs are developed and deployed in a safe manner. Ultimately, we hope this will allow us to realize the benefits of this powerful technology while minimizing the potential for catastrophic outcomes.

Hutiri, W., Papakyriakopoulos, O., & Xiang, A. (2024, June). **Not my voice! A taxonomy of ethical and safety harms of speech generators**. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 359-376). https://dl.acm.org/doi/pdf/10.1145/3531146.3533158

Abstract: The rapid and wide-scale adoption of AI to generate human speech poses a range of significant ethical and safety risks to society that need to be addressed. For example, a growing number of speech generation incidents are associated with swatting attacks in the United States, where anonymous perpetrators create synthetic voices that call police officers to close down schools and hospitals, or to violently gain access to innocent citizens' homes. Incidents like this demonstrate that multimodal generative AI risks and harms do not exist in isolation, but arise from the interactions of multiple stakeholders and technical AI systems. In this paper we analyse speech generation incidents to study how patterns of specific harms arise. We find that specific harms can be categorised according to the exposure of affected individuals, that is to say whether they are a subject of, interact with, suffer due to, or are excluded from speech generation systems. Similarly, specific harms are also a consequence of the motives of the creators and deployers of the systems. Based on these insights we propose a conceptual framework for modelling pathways to ethical and safety harms of AI, which we use to develop a taxonomy of harms of speech generators. Our relational approach captures the complexity of risks and harms in sociotechnical AI systems, and yields a taxonomy that can support appropriate policy interventions and decision making for the responsible development and release of speech generation models.

[UNI only] Messeri, Lisa, & Crockett, Molly J. (2024). **Artificial intelligence and illusions of understanding in scientific research**. *Nature, 627*(8002), 49-58. https://static1.squarespace.com/static/538ca3ade4b090f9ef331978/t/65f071f8fd3e3b478a4f4b86/1710256633821/Messeri%26Crockett_2024_Nature.pdf

Abstract: Scientists are enthusiastically imagining ways in which artificial intelligence (AI) tools might improve research. Why are AI tools so attractive and what are the risks of implementing them across the research pipeline? Here we develop a taxonomy of scientists' visions for AI, observing that their appeal comes from promises to improve productivity and objectivity by overcoming human shortcomings. But proposed AI solutions can also exploit our cognitive limitations, making us vulnerable to illusions of understanding in which we believe we understand more about the world than we actually do. Such illusions obscure the scientific community's ability to see the formation of scientific monocultures, in which some types of methods, questions and viewpoints come to dominate alternative approaches, making science less innovative and more vulnerable to errors. The proliferation of AI tools in science risks introducing a phase of scientific enquiry in which we produce more but understand less. By analyzing the appeal of these tools, we provide a framework for advancing discussions of responsible knowledge production in the age of AI.

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022, June). **The fallacy of AI functionality**. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 959-972). https://dl.acm.org/doi/pdf/10.1145/3531146.3533158

Abstract: Deployed AI systems often do not work. They can be constructed haphazardly, deployed indiscriminately, and promoted deceptively. However, despite this reality, scholars, the press, and policymakers pay too little attention to functionality. This leads to technical and policy solutions focused on "ethical" or value-aligned deployments, often skipping over the prior question of whether a given system functions, or provides any benefits at all. To describe the harms of various types of functionality failures, we analyze a set of case studies to create a taxonomy of known AI functionality issues. We then point to policy and organizational responses that are often overlooked and become more readily available once functionality is drawn into focus. We argue that functionality is a meaningful AI policy challenge, operating as a necessary first step towards protecting affected communities from algorithmic harm.

Slattery, P., Saeri, A. K., Grundy, E. A., Graham, J., Noetel, M., Uuk, R., ... & Thompson, N. (2024). **The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence.** *arXiv preprint arXiv:2408.12622*. https://arxiv.org/abs/2408.12622

Abstract: The risks posed by Artificial Intelligence (AI) are of considerable concern to academics, auditors, policymakers, AI companies, and the public. However, a lack of shared understanding of AI risks can impede our ability to comprehensively discuss, research, and react to them. This paper addresses this gap by creating an AI Risk Repository to serve as a common frame of reference. This comprises a living database of 777 risks extracted from 43 taxonomies, which can be filtered based on two overarching taxonomies and easily accessed, modified, and updated via our website and online spreadsheets. We construct our Repository with a systematic review of taxonomies and other structured classifications of AI risk followed by an expert consultation. We develop our taxonomies of AI risk using a best-fit framework synthesis. Our high-level Causal Taxonomy of AI Risks classifies each risk by its causal factors (1) Entity: Human, AI; (2) Intentionality: Intentional, Unintentional; and (3) Timing: Pre-deployment; Post-deployment. Our mid-level Domain Taxonomy of AI Risks classifies risks into seven AI risk domains: (1) Discrimination & toxicity, (2) Privacy & security, (3) Misinformation, (4) Malicious actors & misuse, (5) Human-computer interaction, (6) Socioeconomic & environmental, and (7) AI system safety, failures, & limitations. These are further divided into 23 subdomains. The AI Risk Repository is, to our knowledge, the first attempt to rigorously curate, analyze, and extract AI risk frameworks into a publicly accessible, comprehensive, extensible, and categorized risk database. This creates a foundation for a more coordinated, coherent, and complete approach to defining, auditing, and managing the risks posed by AI systems.

Policy Papers

Acemoglu, D. (2021). **Harms of AI** (No. w29247). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w29247/w29247.pdf

Abstract: This essay discusses several potential economic, political and social costs of the current path of AI technologies. I argue that if AI continues to be deployed along its current trajectory and remains unregulated, it may produce various social, economic and political harms. These include: damaging competition, consumer privacy and consumer choice; excessively automating work, fueling inequality, inefficiently pushing down wages, and failing to improve worker productivity; and damaging political discourse, democracy's most fundamental lifeblood. Although there is no conclusive evidence suggesting that these costs are imminent or substantial, it may be useful to understand them before they are fully realized and become harder or even impossible to reverse, precisely because of AI's promising and wide-reaching potential. I also suggest that these costs are not inherent to the nature of AI technologies, but are related to how they are being used and developed at the moment - to empower corporations and governments against workers and citizens. As a result, efforts to limit and reverse these costs may need to rely on regulation and policies to redirect AI research. Attempts to contain them just by promoting competition may be insufficient.

Websites

AI Risk Repository. Massachusetts Institute of Technology. https://airisk.mit.edu/

Summary: “The AI Risk Repository has three parts:

- The AI Risk Database captures 700+ risks extracted from 43 existing frameworks, with quotes and page numbers.
- The Causal Taxonomy of AI Risks classifies how, when, and why these risks occur.
- The Domain Taxonomy of AI Risks classifies these risks into seven domains (e.g., “Misinformation”) and 23 subdomains (e.g., ‘False or misleading information’).”

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025, April). **AI 2027.** AI Futures Project. https://ai-2027.com/

"We predict that the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution. We wrote a scenario that represents our best guess about what that might look like. It’s informed by trend extrapolations, wargames, expert feedback, experience at OpenAI, and previous forecasting successes."

"The CEOs of OpenAI, Google DeepMind, and Anthropic have all predicted that AGI will arrive within the next 5 years. Sam Altman has said OpenAI is setting its sights on “superintelligence in the true sense of the word” and the 'glorious future.' What might that look like? We wrote AI 2027 to answer that question. Claims about the future are often frustratingly vague, so we tried to be as concrete and quantitative as possible, even though this means depicting one of many possible futures. We wrote two endings: a “slowdown” and a “race” ending. However, AI 2027 is not a recommendation or exhortation. Our goal is predictive accuracy. We encourage you to debate and counter this scenario. We hope to spark a broad conversation about where we're headed and how to steer toward positive futures. We're planning to give out thousands in prizes to the best alternative scenarios."

Book Chapters/Excerpts

[UNI only] Narayanan, A., & Kapoor, S. (2024). **Cognitive biases lead us astray**. In *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference* (pp. 255-257). Princeton University Press. https://ebookcentral.proquest.com/lib/rodlibrary-ebooks/reader.action?docID=31315362&ppg=255

[UNI only] Narayanan, A., & Kapoor, S. (2024). **News media misleads the public**. In *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference* (pp. 247-251). Princeton University Press. https://ebookcentral.proquest.com/lib/rodlibrary-ebooks/reader.action?docID=31315362&ppg=255

[UNI only] Narayanan, A., & Kapoor, S. (2024). **The reproducibility crisis in AI research**. In *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference* (pp. 241-247). Princeton University Press. https://ebookcentral.proquest.com/lib/rodlibrary-ebooks/reader.action?docID=31315362&ppg=255