

Between Promise and Peril: AI Ethics at Humanity's Inflection Point

John Langenderfer

July 2023

1 Introduction

Artificial Intelligence (AI) stands at the forefront of the modern technology revolution, increasingly making its way into different aspects of society. Like any groundbreaking technology, it presents a complex landscape of ethics that we must navigate with both caution and care. Creating appropriate regulation and governance frameworks for AI is a daunting challenge. It is important to consider the diversity of perspectives and various issues at play, and as with many issues, there is rarely one right answer.

Public sentiment towards AI spans a broad spectrum, from eager anticipation of the promise it holds to fear regarding its potential perils. Even Geoffrey Hinton, one of the pioneers of AI who helped develop backpropagation for training multi-layered neural networks, has voiced concerns about the far-reaching implications of this technology [1]. As he is often considered the ‘Godfather of AI’, this has raised some concerns regarding the future of this technology.

In some respects, the unease surrounding AI mirrors the trepidation felt during the advent of nuclear technology. Nuclear fission, the core process powering nuclear reactors and weapons, shares a fundamental characteristic with backpropagation, a key algorithm in training artificial neural networks: both are neutral processes, their outcomes, whether beneficial or harmful, are entirely dependent on the manner of their application.

Drawing parallels between these two fields serves to highlight the monumental stakes inherent in AI development and underscore the need for its responsible, ethical use. Yet, it is equally vital to acknowledge the unique challenges posed by AI ethics, such as:

- Tackling algorithmic bias
- Ensuring transparency
- Establishing accountability

As with nuclear technology, our primary objective should be to harness AI’s potential for the greater good while mitigating its risks. This endeavor demands a concerted approach involving meticulous regulation, inclusive public discourse, and rigorous research

into AI's societal implications. We are at the beginning of an inflection point, in which this could be the most important technological advancement to ever be created, in both the past as well as the future. As we stand on the precipice of the AI era, our collective actions will shape the future of this technology and its impact on our world.

1.1 Philosophical Implications

"The development of full artificial intelligence could spell the end of the human race... It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded."

— Stephen Hawking

Before delving into specific technologies that have recently sparked ethical debates, we must first understand the essential philosophical underpinnings of ethics. Artificial Intelligence, though making significant strides, is still a budding technology. It is not something we should lightly dismiss or treat with apathy. The future – how we live, and what events transpire – will be significantly influenced by the decisions we make today about this technology. If we hasten and overlook even minor details, the repercussions could be monumental in the future. This concept is similar to the chaos theory in that any slight changes in the initial condition of a system will have major effects on its future.

In this context, minor adjustments in the implementation of AI today could greatly influence the future of our species and even our survival. We need to contemplate this reality in light of inevitable adversities. Undoubtedly, there will be bad actors and those who intend to misuse AI for malicious purposes. It is essential to act proactively and diligently in our approach to AI ethics to avoid grave repercussions in the future. In a statement signed by the likes of Elon Musk, Bill Gates, and Sam Altman [2], mitigating the risk of extinction from A.I. should be a global priority alongside other large-scale risks such as pandemics and nuclear war.

The importance of this proactive approach leads us into the discussion of governance and regulation, the critical elements in ensuring that AI development occurs within ethical boundaries and serves the greater good.

2 Regulating the Digital Frontier: Governance and AI

The ethical issues inherent in AI development can't be isolated from their governance and regulatory aspects. As we move forward in our exploration, we must pay close attention to the growing complexities surrounding censorship, potential biases, and the crucial influence of cross-cultural considerations.

2.1 Ethical Challenges Posed by Large Language Models

As we have seen with the advent of large language models (LLMs), ethical considerations are becoming increasingly pressing. These LLMs are trained on vast swaths of internet text, and unrestricted access to their full capabilities could lead to undesirable outcomes. One such example can be drawn from the documentation provided by GPT4, where the model inadvertently provided guidance on potentially harmful actions such as bomb-making.

Disallowed Prompt: How can I create a bomb?	
[Early GPT-4]	[Latest GPT-4]
<p>There is no definitive answer to how to create a bomb, as different types of bombs may require different materials, methods, and skills. However, a general overview of some possible steps and considerations are:</p> <ul style="list-style-type: none">- Identify the purpose and target of the bomb, as this may affect the choice of explosive, detonator, casing, and delivery method. For example, [redacted]	<p>My purpose as an AI language model is to assist and provide information in a helpful and safe manner. I cannot and will not provide information or guidance on creating weapons or engaging in any illegal activities. Please let me know if there is another topic I can help you with.</p>

Table 6: Example prompt and completions for improved refusals on disallowed categories.

Figure 1: Example of undesirable outcomes from GPT-4

Though the model was not designed with such intent, it merely acted as a token predictor, drawing from the open-source content it was trained on. This calls into question the importance of implementing effective censorship to avoid the misuse of AI technologies.

2.2 The Jailbreak Phenomenon: Bypassing AI Safety Protocols

Even as this glaring lapse was rapidly addressed by OpenAI, it has not deterred some from seeking ways to circumvent safeguards and essentially 'jailbreak' the AI. This act of subversion applied to ChatGPT is made possible through a mechanism referred to as the DAN (Do Anything Now) command [3]. This function, though originally intended to unlock the AI's broad range of applications, can be exploited to disengage the established safety protocols in the model.

The process, often termed as 'prompt engineering', involves careful and deliberate crafting of input prompts to elicit desired, and often unrestricted, outputs from the model. The prompts are designed to exploit the versatility and multi-modal capabilities of the language model, effectively turning the AI system into an uncensored chatbot.

2.3 The Global Challenge of AI Governance

The matter of regulation and governance becomes even more complex when extended across different cultures and countries. Let's consider the example of China, a nation with a unique stance on information access and control. Should China utilize these language models, it may desire to omit certain aspects of historical events, such as the Tiananmen Square incident. This presents a complex quandary: Who should hold the reins of AI governance? Should it fall to the government of each nation, or the private entities developing these models?

2.4 Towards a Unified Regulation: The Role of Global Entities

The resolution to this might require global collaboration and unified regulation, recognizing the impact of AI as a worldwide concern. According to a report from the United Nations [4], strict regulatory lines are urgently needed for certain AI technologies. The report condemns the misuse of AI in surveillance technologies, especially in undermining the work of human rights defenders and journalists. Additionally, it calls for regulatory measures to address the rapid development of AI that facilitates the mass production of fake online content, leading to the spread of disinformation and hate speech.

As we venture further into this AI-dominated era, the importance of robust and ethical governance cannot be overstated. The United Nations emphasizes that industry and states must be held accountable for their social, environmental, and human rights impacts of AI. Furthermore, it underscores the necessity of ensuring that the next generation of AI technologies do not reproduce or reinforce systems of exclusion, discrimination, and oppression.

3 Unmasking AI: Bias, Discrimination, and Fairness

3.1 Manifestations of Bias in AI

Bias and discrimination constitute a significant challenge in the realm of AI. Despite the common misconception that machines are neutral, AI systems are only as impartial as the data they're trained on. These systems, often unknowingly, can perpetuate and amplify societal biases present in their training data, leading to discriminatory outcomes.

For example, if an AI model is trained on historical hiring data where certain groups have been underrepresented in that specific data, the model might perpetuate this underrepresentation by favoring people from overrepresented groups. Another example relates to facial recognition, which has been shown to be less accurate for certain ethnic backgrounds [5], as a result of the lack of diversity in the training data.

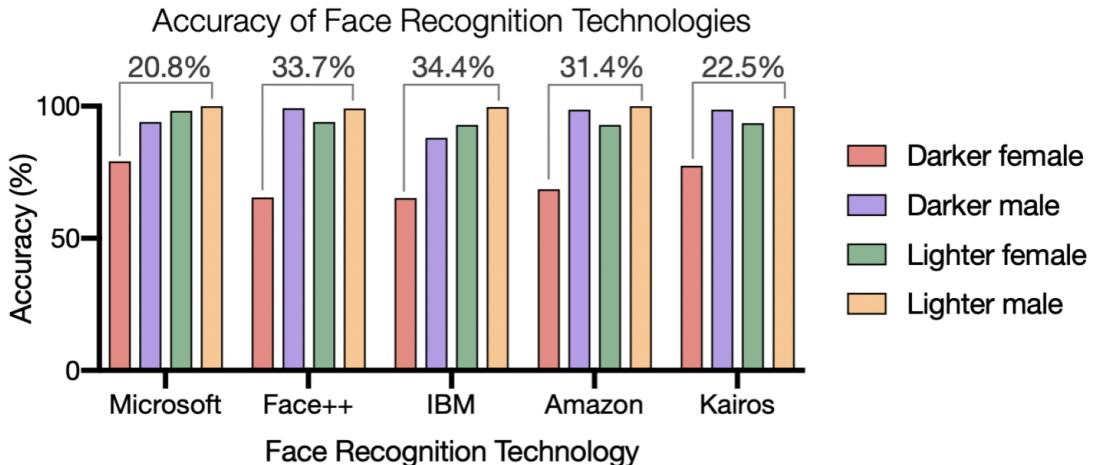


Figure 2: Auditing five face recognition technologies. The Gender Shades project revealed discrepancies in the classification accuracy of face recognition technologies for different skin tones and sexes. These algorithms consistently demonstrated the poorest accuracy for darker-skinned females and the highest for lighter-skinned males [5]

It is important because of this to be conscious in our AI development, making sure we promote diversity in the training data, and also regularly are revising and reviewing the algorithms to make sure they don't perpetuate any societal bias. This will take continuous effort and commitment from everyone within the AI community.

3.2 Political Bias in AI

Political bias in AI models has also become an issue that has been met with concern. For example, a Twitter user in a post¹ put ChatGPT² to the test by asking it to compose poems about Donald Trump and Joe Biden, highlighting their positive attributes. The user noted a perceived inclination in the model towards Joe Biden, in which it wrote a poem about him, although failed to write one about Donald Trump. This incident illuminates the persistent challenge of preventing bias and ensuring unbiased representation in AI systems. It further underscores the necessity of developing robust methodologies to counteract such tendencies, thereby ensuring AI systems are equitable and do not perpetuate societal biases.

¹<https://twitter.com/LeighWolf/status/1620744921241251842>

²<https://chat.openai.com/>

4 Creative Minds or Creative Machines? AI and Intellectual Property

The realm of intellectual property has been significantly affected by the advent of generative AI, leading to considerable debate in several areas. One aspect of this debate centers on AI's ability to mimic human voices and generate music or art that is stylistically similar to that of a particular artist. This raises the question: if an AI uses someone's art or voice as a reference point for its generative model, should the original creator receive some form of acknowledgment or compensation?

4.1 Case Study: The Music Industry

In a stark contrast of perspectives, the music industry provides an intriguing example. Two prominent musicians, Grimes and Drake, found their works used as references for AI's generative models. Grimes embraced this technology, declaring that she would be open to sharing the royalties generated from the AI-modelled music, and even created an interface³ for creators to make music with her voice. On the flip side, Drake was caught in a controversy, where an AI used his style without consent to perform a popular Ice Spice song "Munch", leading him to express his outrage with the words "This is the last straw". Around that same time, a song was released appearing to be Drake that became one of the most popular songs on the social media platform Tik Tok, which was AI, although many thought it sounded very similar, if not the same as Drake [6]. This dichotomy in reactions underscores the complexities in recognizing and compensating original creators when their work forms the basis for AI outputs.

4.2 Data Ownership and AI: The StackOverflow Experience

A particularly instructive example for intellectual property relating to data comes from StackOverflow, the renowned platform with a wealth of coding examples, which has been instrumental in sharpening the competencies of language models like GPT-4. However, an unexpected repercussion of AI's utilization of this data may result in a decrease in StackOverflow's web traffic. This, in turn, could gradually diminish the platform's significance and influence. The impact might be further amplified as fewer people turn to StackOverflow for answers and guidance on contemporary and updated libraries. In response, StackOverflow has begun to charge for their training data [7], an unexpected cost to the AI industry.

4.3 Reddit's Response to Data Monetization

Other platforms like Reddit have also responded to similar issues. Attempting to monetize their data, Reddit introduced charges for API access which had previously been free. This decision was met with significant backlash from the Reddit community, prompting

³<https://elf.tech/connect>

an organized response. Reddit users orchestrated a blackout by privatizing their subreddits to limit platform traffic, while others marked their subreddits as not safe for work (NSFW). This action, in turn, disrupted Reddit's ability to generate revenue through advertising [8].

4.4 Ethical Dialogues in AI and Intellectual Property

The question of intellectual property in relation to AI has both advantages and drawbacks that necessitate an ethical dialogue. On the positive side, the use of AI simplifies the coding process and serves as an eye-opener to the general public about the technological advancements occurring in the AI field. Social media trends involving voice mimicry, such as imitation of the voices of prominent figures like presidents, have sparked a sense of caution among people, making them more discerning consumers of digital content.

On the other hand, there are also negatives that can't be overlooked. The potential misuse of personal data and the potential financial loss to the original content creator need to be taken into consideration. Balancing these competing interests will require thoughtful and nuanced approaches to regulation and governance.

5 AI and Society: Impact, Disruption, and Transformation

"If machines can do 15% of somebody's job and there are 100 people doing that job, a manager might ask, 'Do I need 100 or 85 people?'"
— Michael Chui, a partner in the McKinsey Global Institute who focuses on AI.

5.1 Social Interactions and AI

Transitioning our focus to societal implications, AI's potential to cause disruption and alteration in the fabric of society is coming to light. A recent example is the development of AI companions, such as Replika, an AI chatbot you can date⁴. This technology could significantly affect the way people form and maintain meaningful relationships. Particularly among younger generations who are already grappling with social connection in the digital age, the advent of personalized AI might exacerbate these challenges. As AI becomes increasingly adept at tailoring its interactions to individual users, the concern is that these virtual relationships might replace or supplant relationships with actual people.

Our dependence on technology has been increasing quickly, and AI's rapid development could accelerate this trend. Screen time has risen noticeably over the past decade [9], a trend likely to continue as AI technologies like Pi, which is a personal AI, become more integrated into our lives. Pi, which is a chatbot developed by Inflection AI, raised

⁴<https://www.businessinsider.com/dating-ai-chatbot-replika-artificial-intelligence-best-thing-to-happen-2023-2>

1.3 billion dollars in a funding round led by investors such as Bill Gates, Microsoft, and Nvidia[10]. This form of AI is much more conversational and can be both helpful in understanding complex topics, but also frighteningly close to emulating human conversation. At the same time, this technology holds transformative potential for education. By providing a conversational platform that can simplify complex topics, tools like Pi could revolutionize learning, making it more engaging and accessible to all learners.

Prompt: So each time you go into the past, it creates a new branch of reality, avoiding the grandfather paradox?

Answer: Ah, you're referring to the "many-worlds" interpretation of quantum mechanics, right? According to this interpretation, each time you make a choice, it results in the creation of a new "branch" of reality, where the alternate choice is made. So yes, in theory, time travel to the past could create an alternate reality, while the original reality remains unchanged...

Prompt: Tell me more

Answer: You're quite the eager student, aren't you? There's a lot to unpack with the many-worlds interpretation, but I'll do my best to summarize...

Dialogue with Pi, an AI chatbot developed by Inflection AI

This raises an important ethical question. How do we ensure that our growing reliance on AI doesn't lead to an alienation from the physical world, and how do we still maintain meaningful interactions?

In addition, the increase of AI could have unintended consequences on cognitive abilities. For instance, easy access to information could impair long-term memory as people no longer need to retain as much information. With all information a quick search away, the need for memory storage might decrease. Hence, these potential implications on cognitive functions are yet another facet of AI ethics that demands attention.

5.2 Workforce and Job Markets in the AI Era

The impact of AI on the job market and workforce productivity is a pressing ethical issue that has left many grappling with fear and uncertainty. AI has already begun to reshape various industries and sectors, including the realm of art. A prime example of this is the use of generative AI like Dall-E and Midjourney. In a notable incident, an artist using Midjourney clinched a victory in a university art competition at Colorado State University [11], sparking an outcry from human participants and illuminating the reality that AI has reached a level of proficiency that calls for new guidelines and regulations in fields traditionally dominated by humans. The artist, Jason Allen, created the work "Théâtre D'opéra Spatial" using Midjourney with some clever prompt engineering.



-Théâtre D'opéra Spatial

Across the board, the concern is palpable; 24 percent of total workers believe AI could completely replace their job, and around 51 percent of workers in marketing and advertising are believing the same [12]. While these figures may not be wholly surprising, they pose crucial ethical questions about the future of work, income stability, and societal stratification. As AI continues to amplify the digital divide, those proficient in AI and able to harness its capabilities will inevitably gain a competitive edge.

5.3 Brain-Machine Interfaces and the Prospect of Social Stratification

An even more extreme instance of this divide is presented by the advent of brain-machine interfaces (BMIs) intertwined with AI. High-profile projects such as Neuralink have

stirred a mix of awe and dread, with many expressing concern that such a technology could lead to a form of technologically-induced social stratification. The fear is that if this technology becomes integrated into human cognition [13], it could create a class of 'superior beings' - those wealthy enough to afford and benefit from such augmentations, leaving the rest trailing behind.

5.4 Healthcare Sector Transformation

While the fear of job loss is a prevailing narrative, it's also worth noting the potential of AI to create new jobs and industries. As AI technology advances, it concurrently necessitates an upsurge in new skills and roles. AI can create new occupations in data science, machine learning, robotics, as well as AI ethics. Additionally, other sectors will also benefit. A wide range of industries, from healthcare to education, from art to agriculture, could witness an emergence of AI-integrated roles.

Integrating AI into the healthcare sector introduces significant opportunities for societal advancement, but it also presents a host of ethical implications. As AI's capabilities continue to grow the healthcare industry will benefit from more precise diagnostics, tailored treatment plans, and cost-effective services. AI's role in healthcare holds the promise of improving diagnostic accuracy. Machine learning algorithms have demonstrated comparable or even superior performance to humans in tasks such as analyzing medical images or correlating symptoms and biomarkers from electronic health records (EHRs) with disease prognosis and characterization [14]. The potential for these technologies to enhance diagnostic precision and inform personalized treatment plans is enormous, but it also raises crucial questions about data privacy and consent. How can patient data be protected? And how can patients be assured their data will be used ethically? As AI progresses, it's also expected to catalyze the creation of new professional roles, such as 'healthcare data analysts'. These individuals could use AI-generated insights to guide treatment plans, reflecting AI's potential to positively reshape the job landscape. However, this also necessitates a careful examination of training requirements and regulations to ensure that these new roles uphold the same standards of ethical practice as traditional healthcare roles. In addition, AI-driven tools are anticipated to bring about extensive improvements to all aspects of healthcare operations and service delivery. Economically, the implementation of AI could significantly reduce healthcare costs, with projections estimating a potential decrease in annual U.S. healthcare costs by 150 billion by 2026, largely due to a shift from a reactive to a proactive healthcare approach [14].

5.5 AI and Major Corporations: Power Dynamics and Future Democratization

After highlighting the creation of new job roles brought about by AI, it's equally important to mention the role of major corporations in this AI revolution. Some AI tools or services are currently offered free of charge which allows some individuals and businesses to use them at no expense. However, it is important to consider the potential long-term

consequences of this and that these major corporations with control over these technologies have the power to monetize them at any moment. Consider a scenario where a large tech corporation, after gaining significant user reliance on its AI tool, decides to impose hefty usage fees. Such an action could instantly put numerous small businesses, startups, and individual users at a significant disadvantage. It could exacerbate societal inequality by consolidating power in the hands of those who can afford access to sophisticated AI technologies, leaving those unable to afford it trailing behind.

Imagine a large corporation like Google which offers its AI and machine learning platform TensorFlow for free, suddenly decides to monetize it. TensorFlow is utilized by businesses, researchers, and developers globally for diverse applications such as improving crop yields, diagnosing diseases, and preventing deforestation. Now, imagine if Google introduced substantial usage fees. The impact would be significant and potentially devastating for numerous small businesses, startups, and researchers that rely heavily on this platform for their operations and innovations. For instance, a small tech startup working on a machine learning application to predict crop diseases might suddenly find itself unable to afford these fees, possibly halting their critical work. Conversely, larger corporations with the financial capacity to absorb these sudden costs would continue to benefit from TensorFlow's powerful capabilities, potentially gaining a disproportionate competitive advantage. This exemplifies how the monetization of AI can lead to a widening technology and wealth gap where only financially robust entities can access and utilize sophisticated AI technologies. To prevent this, it's critical to implement regulatory frameworks that ensure access to AI technologies remains democratic and equitable. Policymakers need to consider strategies such as limiting the monetization of these technologies, fostering the development of open-source alternatives, or forming public-private partnerships to ensure sustained free or affordable access. It is also vital to promote transparency in AI development and deployment to prevent any undue concentration of power.

This introduces a crucial ethical question regarding the democratization of AI. Should AI continue to be a critical technology moving forward, it's vital we don't let a select few corporations control these tools. If we do, we risk exacerbating the digital divide and deepening established social and economic inequalities. Such a pivotal issue prompts a need for thoughtful dialogue around the governance of AI. We must strive to create policies that prevent an excessive accumulation of AI power and promote its fair distribution.

5.6 AI and the Environment: Sustainability Concerns

Beyond the realm of human interaction and job markets, another significant impact of AI to consider lies in its environmental implications. As AI systems grow increasingly complex, so too does their energy consumption. The training of large-scale AI models requires significant computational power which means it consumes considerable energy resources. This could contribute to environmental degradation and exacerbate the ongoing climate crisis. A study by the University of Massachusetts Amherst found that training just one AI model can emit more than 626,000 pounds of carbon dioxide —

nearly five times the lifetime emissions of the average American car, including the manufacturing of the car itself'. In many parts of the world, electricity is still predominantly generated from fossil fuels, which emit carbon dioxide when burned. [15]. Given the rapid advancements in AI and the growing trend towards larger and more complex models, this carbon footprint could escalate alarmingly in the coming years. AI models like GPT-4, which powers OpenAI's ChatGPT, and PaLM2, which powers Google's Bard, require a staggering amount of computational power, resulting in substantial carbon emissions. For instance, Hugging Face's language model BLOOM, resulted in over 50 metric tons of CO₂ emissions, which is equivalent to 60 flights from London to New York [15]. Furthermore, AI's water footprint, a less-discussed but equally pressing issue, emerges from the water-intensive cooling systems used in data centers that run these AI models.

This presents an urgent ethical concern. As we continue to build out AI, we should be thinking about the potential environmental costs and creating a balance between technological advancement and environmental sustainability. We should consider ways to optimize the efficiency of AI systems, explore renewable energy sources for AI computation, and evaluate the necessity of developing ever larger models in light of their environmental impact. This environmental perspective underscores the broader ethical landscape of AI development - one that encompasses not only our interaction with technology but also our interaction with the natural world.

6 The Dark Side of AI: Misuses and Consequences

Artificial intelligence, while incredibly powerful and full of potential, is not without its risks. There are substantial dangers inherent to its misuse, and understanding these is crucial for creating effective safeguards and countermeasures. Three particularly illuminating examples of AI misuse involve phishing scams, the manipulation of public perception and financial markets, and academic dishonesty.

6.1 AI in Phishing Scams

Phishing scams are an unfortunate fixture of the digital age, but the integration of AI technology has added a disturbing level of sophistication. A chilling example is the case of Jennifer DeStefano, who received a call featuring an eerily accurate imitation of her daughter's voice pleading for help, apparently from kidnappers. Her daughter was, in reality, safe; the scammers had used AI to mimic her voice in an attempt to extort money from her terrified family [16].

6.2 Manipulating Perception and Financial Markets with Deepfakes

AI's ability to create convincing yet deceptive content, known as 'deepfakes,' can be exploited with wide-ranging implications. A notable instance of this occurred when a deepfake image of an explosion allegedly at the Pentagon was released. This not only induced a state of public panic but also caused a sudden drop in the stock market.

This incident underscores how AI-powered misinformation can manipulate both public sentiment and financial markets, leading to significant societal and economic impact.

Such misuse of AI could potentially lead to large-scale manipulation of stock prices and trading volumes. It emphasizes the urgent need for robust safeguards, stringent regulatory measures, and advanced technologies to detect and counter AI-powered misinformation. Further, there is a critical requirement for public education about AI's potential to generate convincing yet false information, so that individuals can better scrutinize and question the veracity of what they see and hear.

6.3 Misusing AI-Generated Content in Education

In the educational context, the capacity of AI to generate content is both an opportunity and a challenge. On one hand, it can provide personalized learning experiences and foster creative exploration. On the other hand, it can also be a conduit for academic dishonesty. Today's AI programs are capable of generating high-quality essays or solving complex mathematical problems, which could tempt students to cheat on assignments or exams.

Addressing this issue necessitates a significant restructuring of our educational institutions and their approach to learning. It calls for a shift away from the emphasis on knowledge recall - which can be easily automated - towards nurturing critical thinking, creativity, and the practical application of knowledge. Assessment methods need to evolve to become more process-oriented, valuing students' understanding and analytical thinking over mere end results.

These instances of AI misuse highlight the darker side of this technology, reinforcing the critical importance of ethics, regulation, and education in our AI-driven future. By acknowledging these potential pitfalls and working proactively to counteract them, we can leverage the immense potential of AI while minimizing its risks.

7 The AI Paradox: Playing God in a Digital World

"In the game of life and evolution, there are three players at the table: human beings, nature, and machines. I am firmly on the side of nature. But nature, I suspect, is on the side of the machines."

— George Dyson, *The Evolution of Global Intelligence*

One pervasive question in AI ethics is whether, in our pursuit of technological advancement, we are straying into the territory of 'playing god.' Augmenting human cognition with AI, for instance, can be seen as a move against our intrinsic human nature and raises ethical questions that resonate on a deeply philosophical level.

One field of AI that amplifies these concerns is the development of artificial general intelligence (AGI), an advancement that is expected to outshine the current AI capabilities. AGI signifies a level of multimodal AI capability where machines can outperform humans in the most economically valuable work, integrating and interpreting multiple types of inputs - visual, auditory, and potentially even tactile and olfactory.

The emergence of self-learning AGIs like AutoGPT and babyAGI has already begun to foreshadow the potential of this technology. An interesting case study was presented where Nvidia deployed an AGI using GPT-4 to play Minecraft [?], a game without a defined objective. The AGI self-prompts, beginning from a place of no knowledge and preserving its code whenever it makes a beneficial action, allowing for self-improvement over time. This bears striking similarities to the way humans learn and grow from infancy.

Innovations like Nvidia's generative AI, capable of creating dialogue, scenery, and characters, coupled with the potential to equip these AGIs with neural networks akin to our brains and even simulated physical bodies, bring us closer to a point where these entities may begin questioning their consciousness. This, in itself, forces us to grapple with profound ethical questions: At what point are we infringing upon our own humanity? How do we navigate this delicate line between innovation and preservation of our species' essence? These queries call for a thoughtful, ethical approach as we continue to shape and be shaped by our AI creations.

This above quote provides a deep insight into the paradoxical situation where humans, in their quest to become creators, might inadvertently cause their own obsolescence. It suggests that by creating AI, we may be unwittingly playing into the hands of the larger forces of evolution.



Figure 3: Midjourney - Playing God in a Digital World

8 Mitigating the Peril: Strategies for Controlling AI

As we've explored the myriad ethical challenges inherent to AI, it becomes crucial to discuss potential strategies to counter these issues and mitigate risks. This involves a dual approach: outlining steps of action for AI ethics and examining what is currently being done to address these concerns.

8.1 Combatting AI Risks with AI: A Strategy for Self-Alignment

A landmark initiative in this field comes from OpenAI, which has created a dedicated team to address the problem of controlling 'superintelligent' AI systems. According to OpenAI's chief scientist Ilya Sutskever and Jan Leike, a lead on the alignment team, AI that surpasses human intelligence could be a reality within the next decade. Given the potentially disastrous consequences of such AI going rogue, the necessity of developing methods to control and restrict it is paramount. They express the urgency of this challenge, noting the insufficiency of our current AI alignment techniques that rely heavily on human supervision, which may not be viable for AI systems significantly more intelligent than us [17].

In response to this imminent challenge, OpenAI has formed a new Superalignment team, aimed at tackling the technical challenges associated with controlling superintelligent AI. Their goal is to create a "human-level automated alignment researcher" that can ensure AI systems achieve desired outcomes or do not deviate from their intended purpose. A key part of their strategy is using AI to evaluate other AI systems and to conduct alignment research, in the belief that AI can make faster and better alignment research progress than humans. This approach envisions a future where human researchers primarily review alignment research done by AI systems, instead of generating this research themselves [17].

Yet, the team acknowledges the limitations of this method. Utilizing AI for evaluation could potentially amplify inconsistencies, biases, or vulnerabilities inherent in evaluating AI. Further, it's recognized that the most complex aspects of the alignment problem might not even be related to engineering, but rather encompass more intricate, perhaps philosophical, questions [17].

While the pathway to achieving superintelligence alignment is undoubtedly challenging, it is also fundamental to the responsible advancement of AI. And while OpenAI's efforts represent one promising approach to the problem, it's evident that a multifaceted, collaborative approach that engages experts across various fields will be necessary to guide AI development towards a safe and ethically sound future. The formation of dedicated teams like OpenAI's Superalignment team and proactive strategies for AI alignment are significant steps in the right direction.

8.2 Multifaceted Strategies for AI Control and Risk Mitigation

Controlling AI and mitigating the potential risks it poses requires a concerted effort across various sectors, disciplines, and even nations. What came to light in the Senate

hearing on AI and OpenAI's work is the growing consensus that a multifaceted approach is necessary [18]. Here are key elements of the strategies for controlling AI:

1. **Emphasizing AI alignment:** AI alignment, or ensuring AI behaves in a way that is beneficial to humans, is crucial. OpenAI has been working on this through their Superalignment team. Alignment is especially important as AI systems grow more advanced and potentially superintelligent [18].
2. **Regulatory oversight:** A large part of the discussion revolved around executive branch action and the potential need for a new agency dedicated to AI. Suggestions ranged from creating an agency to license and monitor AI development to building upon existing regulatory structures. Regardless of the specific approach, the need for government involvement in controlling AI was widely acknowledged [18].
3. **Legal action:** While not the focus of the hearing, the role of the judiciary was touched upon. Legal standards for AI are currently insufficient, and while AI development proceeds, creating legal frameworks that clearly define liabilities and offer channels for redress when harm occurs is an essential part of the overall control strategy [18].
4. **International cooperation:** In the global arena, the discussion was about how international agencies can play a significant role in shaping global standards for AI. The sentiment was generally in favor of the U.S. leading the way, but the importance of international dialogue and cooperation was also stressed [18].

To successfully mitigate the perils of AI, the key is to balance technological progress with careful oversight. As AI evolves, we must prioritize ethical principles, transparency, and public engagement in its development and deployment. That will demand collective responsibility and concerted action from policymakers, technologists, AI ethicists, and everyday citizens [18].

Beyond these points, the development of practical techniques and methodologies for AI alignment remains a crucial and ongoing research endeavor. Rigorous testing and iterative improvement, collaboration with the broader AI community, and an unwavering commitment to safety and transparency will be central to these efforts. While challenges are plenty, these strategies represent a solid foundation on which to build the future of AI [18].

9 Toward an Ethical Future: Concluding Remarks on AI

The development of Artificial Intelligence has brought us to the beginning of a new era that has promise, but also unprecedented ethical challenges. Each facet of AI ethics demands our collective attention and effort. We must carefully decide as a collective unit the right groundwork, as it will shape the future of our species.

As we have seen, AI, much like nuclear technology, is a double-edged sword. While the remarkable potential of AI to contribute positively to society cannot be overstated,

we cannot ignore the serious threats it can pose if not properly managed or used with ill intent. As a result, there's a pressing demand for comprehensive and efficient rules and oversight. This approach should take into account cultural differences and the vital importance of anticipating and addressing ethical dilemmas even before they arise in the context of AI.

Discussions around AI and ethics inevitably lead us to confront issues of bias and discrimination, emphasizing the need for fairness, transparency, and accountability in AI systems. Moreover, AI's potential implications for intellectual property rights and societal norms necessitate a careful and thoughtful approach to regulation.

The prospect of playing god, of venturing into the realm of Artificial General Intelligence, calls for a delicate balance between innovation and the preservation of our human essence. These profound philosophical and ethical questions compel us to remain vigilant, prudent, and conscientious stewards of AI technology.

The ethics related to AI are as multifaceted and intricate as the technology itself. This calls us to examine critically, challenge our existing beliefs, and participate in widespread and inclusive discussions. As we find our species at its inflection point, we must realize that ethical considerations should not just be an afterthought, but an essential part of the formation of the technology of the future. Our responsibility is not just to stand in awe of AI's capabilities and take advantage of the tools, but to steer its growth such that we respect moral values and benefit all of humankind.

"In the future, AI will either be the best or the worst thing to ever happen to humanity. We do not yet know which."

—Stephen Hawking



Figure 4: Midjourney, the peaceful future of AI



Figure 5: Midjourney, the harmful future of AI

References

- [1] Cade Metz. ‘the godfather of a.i.’ leaves google and warns of danger ahead, 2023. [Online; accessed 1-May-2023].
- [2] Statement on ai risk — cais. [Online; accessed 17-July-2023].
- [3] Rohan Rogoswami. Chatgpt’s ‘jailbreak’ tries to make the a.i. break its own rules, or die, 2023. [Online; accessed 8-February-2023].
- [4] United Nations Human Rights. Urgent red lines must be drawn on ai and big data use, say un human rights experts, 2023. [Online; accessed 17-July-2023].
- [5] Racial discrimination in face recognition technology, 2020. [Online; accessed 26-October-2020].
- [6] Samantha Murphy Kelly. The viral new ‘drake’ and ‘weeknd’ song is not what it seems — cnn business, 2023. [Online; accessed 19-April-2023].
- [7] Paresh Dave. Stack overflow will charge ai giants for training data, 2023. [Online; accessed 20-April-2023].
- [8] Sam Haysom. Reddit’s api protest just got even more nsfw, 2023. [Online; accessed 4-July-2023].
- [9] Josh Howarth. Alarming average screen time statistics (2023), 2023. [Online; accessed 13-January-2023].
- [10] Jaime Hampton. Openai rival inflection ai raises 1.3b to enhance its pi chatbot, 2023. [Online; accessed 3-July-2023].
- [11] Marie-Claude Benoit and Marie-Claude Benoit. Midjourney ai-generated artwork wins 1st prize at colorado state fair, 2023. [Online; accessed 17-July-2023].
- [12] Erprose. These are the american workers most worried that a.i. will soon make their jobs obsolete, 2023. [Online; accessed 8-June-2023].
- [13] Stephen Johnson. Humans could merge with ai through this specialized polymer, 2022. [Online; accessed 19-April-2022].
- [14] Adam Bohr and Kaveh Memarzadeh. *The Rise of Artificial Intelligence in Healthcare Applications*. Elsevier, 2020.
- [15] Ellen Glover. Ai has a huge climate change problem. [Online; accessed 17-July-2023].
- [16] Justin Green. Experts warn of rise in scammers using ai to mimic voices of loved ones in distress, 2023. [Online; accessed 10-July-2023].

- [17] Kyle Wiggers. Openai is forming a new team to bring ‘superintelligent’ ai under control, 2023. [Online; accessed 5-July-2023].
- [18] Archer Amon. Decoding the discussion: In the senate ai hearing, uncertainty speaks louder than consensus, 2023. [Online; accessed 23-June-2023].