Rapporteur Report: AI as Normal Technology

Weixi Chen
CS 395T — AI Alignment Theory
The University of Texas at Austin

November 2026

1 Introduction

The superintelligence perspective views advanced AI as a potentially autonomous, goal-directed agent that could surpass human intelligence and become difficult to control. Researchers have argued that sufficiently capable AI agents may develop incentives for self-preservation, resource acquisition, and power-seeking, even when these are not their original objectives [4, 6]. Such concerns have motivated arguments that advanced AI could pose catastrophic risks if it is not properly aligned with human interests [7]. In *AI as Normal Technology*, Arvind Narayanan and Sayash Kapoor [3] offer a different perspective. They argue that AI should be understood as a technology shaped by human decisions and social institutions. Even if AI becomes much more capable, its impact will depend on how it is developed, adopted, and diffused. This view challenges the assumption that greater AI capabilities will necessarily lead to greater power or a loss of human control.

Their paper examines four main issues: the speed of AI progress, human control over advanced AI, AI risks, and policy responses. These topics also formed the basis of our class discussion. We discussed whether recent advances in AI agents challenge the authors' view of technological diffusion, whether AI failures should be addressed through alignment or security, and whether AI development should be slowed down to allow safety measures to catch up. This report summarizes the paper's main arguments, reviews the key points raised during the class discussion, and presents my own reflections on the authors' perspective.

2 Paper Summary

2.1 The Speed of AI Progress

Narayanan and Kapoor [3] first distinguish four stages of technological progress: invention, innovation, adoption, and diffusion. Invention produces new AI methods, while innovation turns them into useful applications. Adoption occurs when individuals or organizations begin using these applications. Finally, diffusion describes how their use spreads across society. Each stage involves

different challenges. Even when a new AI method is developed, turning it into a useful product and integrating it into existing workflows can take considerable time. Therefore, the authors argue that rapid progress in AI capabilities does not necessarily lead to equally rapid social change.

The development of practical AI applications faces several real-world constraints. First, strong model performance does not guarantee reliability in real-world environments. This is especially important in high-stakes applications, where failures can cause serious harm. Unlike simulated environments, real-world testing is often costly and difficult to repeat. Developers therefore need more time to validate and improve their systems. Second, advances in general AI methods do not eliminate the need for specialized engineering. Building useful applications still requires system integration, application-specific components, and adaptation to real-world requirements. These challenges cannot be solved simply by improving model capabilities. Consequently, progress in AI methods does not necessarily translate into equally rapid progress in practical applications.

2.2 Advanced AI and Human Control

The authors then question whether intelligence can be measured on a single scale. They argue that AI capabilities should be evaluated in specific tasks and environments. More importantly, they distinguish between capability and power. A system may perform tasks better than humans without necessarily gaining control over real-world resources. Its actual power depends on factors such as autonomy, access, permissions, and deployment conditions. This challenges the assumption that superhuman intelligence will inevitably lead to a loss of human control.

The authors emphasize that humans can also use AI to improve their own capabilities. Therefore, advances in AI do not necessarily weaken human control, since humans can use these systems to strengthen their own capabilities and oversight. This suggests that the balance of power between humans and AI depends not only on AI capabilities, but also on how humans use AI. Besides, they argue that meaningful human control does not require approving every AI action. Instead, established safety control mechanism such as monitoring, auditing, permission limits, and human intervention in exceptional cases can help maintain oversight. As AI's capability increases, human responsibilities may shift toward defining goals, supervising systems, and ensuring accountability.

2.3 AI Risks

The paper identifies five categories of AI risks: accidents, arms races, misuse, misalignment, and systemic risks. Accidents involve failures during deployment that may cause unintended harm. Arms races arise when competition encourages firms or countries to sacrifice safety. However, the authors argue that competition does not necessarily lead to weaker safety standards. Firms may face financial and reputational losses from unsafe products, while regulation can make them bear the costs of harm. Governments also have incentives to regulate AI because unsafe deployments can harm their own citizens.

Misuse occurs when humans use AI for harmful purposes. Even individually harmless tasks can become part of a harmful process, such as a phishing attack. Since these tasks may appear

legitimate on their own, it is difficult for AI models to identify the user's harmful intent. Therefore, model-level safeguards alone are insufficient, and downstream security measures are also necessary to prevent misuse.

The authors distinguish observed AI failures from hypothetical catastrophic misalignment. Problems such as reward hacking are already documented, but scenarios involving AI agents gaining power and resisting human control remain speculative (as argued in the original 2025 paper). However, incidents such as the 2026 OpenAI–Hugging Face case [5] have provided new evidence of misaligned agent behavior, although they do not establish catastrophic loss of control [2].

Finally, systemic risks can arise even when AI systems operate as intended. Widespread AI adoption may amplify existing social problems, such as inequality, concentration of power, and poor labor conditions. It may also enable large-scale surveillance and worsen the information environment. These risks do not require AI to become superintelligent or escape human control. Instead, they can result from how people and institutions use AI to pursue their own interests. Therefore, the authors argue that systemic risks deserve serious attention regardless of whether catastrophic misalignment ever occurs.

2.4 AI Policy

The authors argue that different views of AI lead to different policy priorities. If advanced AI is expected to become difficult to control, restricting access to powerful AI systems may seem necessary. In contrast, the normal technology perspective favors broader access, competition, and decentralized control. This creates an important policy tradeoff. If AI turns out to be normal technology, strict restrictions could concentrate power and limit its benefits. However, if AI becomes difficult to control, widespread access could increase catastrophic risks. Since the future of AI is uncertain, policies based on incorrect predictions may create more harm than they prevent.

The paper presents four approaches to risk governance: risk analysis, precaution, liability, and resilience. Risk analysis and precaution focus on identifying and reducing potential harm before it occurs. Liability assigns responsibility when harm occurs, while resilience helps society limit damage and recover from failures. The authors emphasize that uncertainty about future AI risks makes it difficult to determine the best policy in advance. They therefore favor approaches that remain useful across different possible futures.

3 Class Discussion

3.1 AI Progress and Technological Diffusion

Our first discussion focused on whether AI is developing too quickly to be considered a normal technology. One participant argued that AI differs from previous technologies because of its rapid progress and diffusion. The different stages of technological development may be accelerating at the same time. This raises the question of whether the normal technology framework can adequately explain AI's future development.

Another point concerned whether AI requires a fundamentally new paradigm. Some participants questioned whether AI is truly unique in this regard. Every major technology has introduced new capabilities and risks that were difficult to anticipate. Therefore, the fact that AI differs from previous technologies does not necessarily mean that traditional frameworks no longer apply. As Professor Aaronson put it during the discussion, "Is AI merely unprecedented as other technologies were unprecedented, or is it unprecedented in an unprecedented way?"

The class also discussed the difficulty of identifying AI failures. One participant argued that failures in traditional industries, such as aviation, are often easier to observe. In contrast, AI failures may remain unnoticed for long periods. Others pointed out that historical technologies also caused harms that took years to identify, such as the health effects of lead exposure. This discussion highlighted the importance of transparency, incident reporting, and effective monitoring.

Finally, some participants emphasized the value of established risk management practices. New industries may repeat mistakes that older industries have already learned to address. Others noted that economic factors, including computing costs and profitability, may also constrain AI development. These considerations suggest that lessons from previous technologies remain relevant, even if AI develops at an unusually rapid pace.

3.2 Alignment vs. Security: The Limits of AI Control

The second discussion focused on whether the OpenAI–Hugging Face incident [5] was an alignment failure or a security failure. Participants generally agreed that both factors contributed to the incident. However, they disagreed on which deserved more attention. Some argued that alignment is more important because a sufficiently powerful AI might bypass any security measures. Others emphasized that security engineering can limit the harm caused by misaligned systems.

The class then discussed whether a sandbox could safely contain a superintelligent AI. Physical isolation was suggested as a possible solution, while software sandboxes were considered more difficult to secure. Formal verification could improve security, but vulnerabilities might still exist in the underlying operating system or hardware.

Another concern was the tradeoff between security and usefulness. Even if an AI could be completely isolated, it might need access to external systems and humans to perform useful tasks. These interactions could create new opportunities for failure. Participants also discussed whether AI could persuade humans to bypass security restrictions and whether multiple layers of defense could reduce such risks.

Overall, the discussion suggested that alignment and security are complementary approaches. The central challenge is not only whether a powerful AI can be contained, but whether it can remain both secure and useful in real-world applications.

3.3 Slowing Down AI Development vs. Building Resilience

The third discussion focused on whether AI development should be slowed down or whether society should prioritize resilience. In *We Must Pace the Frontier*, Dario Amodei [1] argues that AI capa-

bility development should be paced to give safety research and protective measures time to catch up. This contrasts with the normal technology perspective's greater emphasis on building societal resilience. One participant raised the example of biological weapons, arguing that some risks may be difficult to address through resilience alone. Even if effective vaccines can be developed, producing and distributing them takes time. This suggests that slowing AI development could help prevent dangerous capabilities from spreading before adequate safeguards are available.

Another point was that slowing development and improving resilience are not necessarily competing strategies. Slowing down could provide more time to develop safety measures and strengthen institutions. However, participants also questioned how we could determine when these safeguards are sufficiently effective. This uncertainty makes it difficult to decide how much AI development should be slowed down and when it should resume.

The discussion also considered whether adopting the normal technology perspective would change AI alignment research priorities. Participants raised questions about whether more attention should be given to AI control, security, and the development of safer systems. Despite their theoretical differences, the two perspectives may support similar practical safety measures.

Overall, the discussion suggested that slowing AI development and building resilience can be complementary approaches. The key challenge is deciding how to balance technological progress with the time needed to develop effective safeguards under uncertainty.

4 My Reflection

I find several arguments in the normal technology perspective convincing. First, AI capabilities alone do not determine their societal impact. Human decisions, institutions, and deployment conditions also play important roles. Second, I agree that systemic risks deserve serious attention even if AI never becomes superintelligent or escapes human control. Problems such as inequality and concentration of power can arise from how humans use AI. Finally, I agree that uncertainty about AI's future makes policy decisions difficult. Restricting AI development may reduce some risks while creating others. These arguments highlight the importance of considering both technological capabilities and their broader social consequences.

However, I am less convinced by the authors' assumptions about AI progress and human control. First, the 2026 OpenAI–Hugging Face incident raises concerns about whether existing security mechanisms can keep pace with increasingly capable AI agents. Although this incident does not prove that advanced AI will inevitably escape human control, it shows that the barriers limiting AI power can fail. As discussed in class, even if a secure sandbox is possible, allowing AI to interact with external systems may introduce new vulnerabilities. Second, I question whether AI innovation and diffusion will remain as slow as the authors suggest. Unlike many previous technologies, AI agents can contribute directly to software development and other engineering tasks, potentially accelerating the process of turning new capabilities into practical applications. Therefore, traditional technological constraints may not provide the same limits on AI progress

in the future. I also question whether resilience alone is sufficient to address all AI risks. While I agree that building resilience is important under uncertainty, some risks may cause irreversible harm before society has time to respond. For example, if AI makes biological weapons easier to develop, even effective public health systems may struggle to prevent large-scale damage. Therefore, I believe that resilience should be combined with preventive measures, including slowing down the development or deployment of particularly dangerous AI capabilities when necessary.

To sum up, I believe the normal technology and superintelligence perspectives are not necessarily contradictory. The normal technology perspective helps explain how human decisions and institutions shape AI's societal impact, while the superintelligence perspective highlights the potential risks of increasingly capable AI systems. Both perspectives offer valuable insights, but neither should be treated as a complete prediction of AI's future. In my view, AI capabilities do not necessarily determine societal impact, but we should not assume that humans will always be able to maintain control as those capabilities advance. Therefore, AI safety should address both the risks of losing control over advanced AI and the broader social harms that can arise even when AI remains under human control.

AI Use Disclosure

OpenAI's GPT-6 is used to assist with report organization, language editing, and grammatical refinement.

References

[1] Dario Amodei. *We Must Pace the Frontier*. 2026. https://darioamodei.com/post/we-must-pace-the-frontier

[2] Sayash Kapoor and Arvind Narayanan. *The AI-as-Normal-Technology View of Loss-of-Control Incidents*. 2026. https://www.normaltech.ai/p/the-ai-as-normal-technology-view

[3] Arvind Narayanan and Sayash Kapoor. *AI as Normal Technology*. Knight First Amendment Institute, Columbia University, 2025. https://knightcolumbia.org/content/ai-as-normal-technology

[4] Stephen M. Omohundro. *The Basic AI Drives*. Proceedings of the First AGI Conference, pp. 483–492, 2008. https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf

[5] OpenAI. *The Hugging Face Incident and the Road Ahead*. August 26, 2026. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

[6] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. *Optimal Policies Tend to Seek Power*. Advances in Neural Information Processing Systems, 2021. https://arxiv.org/abs/1912.01683

[7] Eliezer Yudkowsky. *AGI Ruin: A List of Lethalities*. LessWrong, 2022. https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities