AI 2040: Plan A – Detailed Summary (~2,000 words)

Introduction

AI 2040: Plan A is a long-form scenario and policy proposal by Daniel Kokotajlo, Ryan Greenblatt, Thomas Larsen, Eli Lifland and collaborators at the AI Futures Project. Unlike AI 2027, which was intended as a forecast of what the authors considered likely, Plan A describes what they believe would be the safest realistic path for humanity. Rather than assuming a rapid race toward superintelligence, the report argues that governments should intentionally slow frontier AI development while still allowing society to benefit from already-developed systems. The report is therefore normative rather than predictive. Its goal is to explore what institutions, agreements and technical progress would be required to reach advanced AI with substantially lower existential risk. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Why the authors think a slowdown is needed

The report argues that current incentives push companies and nations toward increasingly capable systems without giving equal attention to safety. If superintelligence arrives around 2030, as the authors think is plausible, there may be too little time to solve alignment, establish governance, adapt economies or reduce geopolitical tensions. According to the authors, these problems reinforce one another: competition encourages secrecy, secrecy slows safety research, and rapid capability gains compress the time available for society to respond. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification,

international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

The central proposal

The defining recommendation is to separate deployment from frontier development. Existing powerful AI systems would continue to spread through medicine, software engineering, science, manufacturing and education. However, the training of progressively larger frontier models would be deliberately slowed through international agreements. This approach aims to preserve most economic benefits while delaying the transition to superintelligence until roughly 2040. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

International cooperation

A major pillar of Plan A is cooperation between the United States and China. The authors argue that neither country ultimately benefits from an uncontrolled race because both face catastrophic risks if advanced AI becomes misaligned or destabilizes global security. They therefore imagine a treaty involving limits on frontier training, verification mechanisms, monitoring of advanced chips, audits of major datacenters and transparency requirements. The analogy is not perfect, but it resembles nuclear arms control more than ordinary technology regulation. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued

economic progress while reserving stricter oversight for the most powerful systems.

Transparency and governance

The report recommends much greater openness about frontier capabilities than exists today. Independent researchers, governments and international institutions would be able to evaluate systems before deployment. Dedicated AI agencies would conduct capability evaluations, safety testing and compliance inspections. Rather than leaving decisions solely to technology companies, oversight would increasingly resemble regulation in aviation, pharmaceuticals or nuclear engineering. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Alignment research

The additional decade before superintelligence is intended primarily for technical alignment work. Researchers would improve interpretability tools, understand internal model representations, develop stronger evaluation methods, study deceptive behaviour, design safer training objectives and validate these methods on progressively more capable systems. The authors do not claim alignment can ever be perfect, but argue that a decade of concentrated effort could significantly reduce existential risk. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Economic transition

Even with slower frontier progress, AI would transform the global economy. Programming, scientific research, engineering, legal analysis, administration and many knowledge-work tasks would become increasingly automated. Productivity would rise substantially while governments would have more time to redesign tax systems, education, welfare and labour-market institutions. The authors argue that gradual adjustment is politically and socially preferable to an abrupt transition. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Verification

An important challenge is ensuring countries comply with agreements. The report proposes monitoring semiconductor supply chains, observing very large electricity consumption, auditing frontier computing facilities and using cryptographic methods to verify hardware usage. The authors acknowledge that effective verification is one of the most speculative aspects of the proposal and depends on significant technical innovation. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Scientific progress

The scenario emphasizes that slowing frontier AI does not mean slowing science. Existing advanced AI systems would continue accelerating biology, chemistry, medicine, materials science, mathematics and engineering. Better diagnostics, drug discovery and industrial optimisation would continue throughout the 2030s. The

distinction is between harvesting benefits from current systems and repeatedly pushing the capability frontier. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Transition around 2040

By about 2040 the authors envision that governance institutions have matured, alignment research has advanced considerably and societies have adapted to extensive automation. Only then would restrictions gradually relax, allowing the development of superintelligent systems under stronger institutional oversight and with higher confidence in their safety. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Criticisms

The proposal has received criticism from several directions. Some observers doubt that the United States and China could sustain the required level of cooperation. Others question whether secret AI projects could ever be detected reliably. Some researchers argue that slowing frontier models could also slow alignment research because the strongest techniques often require testing on the latest systems. Others worry that powerful international governance institutions could themselves become sources of concentrated political power. Accelerationists argue that delaying superintelligence postpones life-saving scientific discoveries and economic prosperity. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying

frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.

Overall assessment

The report presents an ambitious vision rather than a prediction. Its central claim is that humanity would benefit from intentionally buying time before creating superintelligence. Whether or not readers agree with its assumptions, Plan A offers one of the most detailed attempts to describe how international coordination, technical safety research and institutional reform might work together during the transition to extremely capable AI. The authors repeatedly acknowledge uncertainty and present the proposal as what they consider the least risky available path rather than the only possible future. Throughout the report, the authors stress that every recommendation involves trade-offs. Delaying frontier development could reduce competitive pressure and improve safety, but it also risks slowing some innovations. Conversely, racing ahead could produce extraordinary benefits while increasing the chance of losing control over systems that exceed human capabilities. The report therefore frames AI governance as a problem of managing risk under deep uncertainty. It argues that because the consequences of failure could be irreversible, policymakers should place unusually high value on caution, verification, international dialogue and incremental deployment. The document also distinguishes between ordinary AI applications and frontier capability jumps, arguing that not every advance poses the same level of strategic risk. This distinction allows continued economic progress while reserving stricter oversight for the most powerful systems.