Transformative AI and Scenario Planning for AI X-risk

15

by **Elliot_Mckernon, Justin Bullock** 22nd Mar 2024

Convergence Analysis (org) | Transformative AI | AI | Frontpage

This post is part of a series by the AI Clarity team at Convergence Analysis. In our previous post, Corin Katzke reviewed methods for applying scenario planning methods to AI existential risk strategy. In this post, we want to provide the motivation for our focus on transformative AI.

Overview

We argue that "Transformative AI" (TAI) is a useful key milestone to consider for *AI scenario analysis*; it places the focus on the socio-technical *impact* of AI and is both widely used and well-defined within the existing AI literature. We briefly explore the literature and provide a definition of TAI. From here we examine TAI as a revolutionary, general purpose technology that could likely be achieved with "competent" AGI. We highlight the use of the Task Automation Benchmark as a common indicator of TAI. Finally, we note that there are significant uncertainties on the time differences between when TAI is *created*, when it is *deployed*, and when it *transforms society*.

Introduction

The development of artificial intelligence has been accelerating, and in the last few years we've seen a surge of shockingly powerful AI tools that can outperform the most capable humans at many tasks, including cutting-edge scientific challenges such as predicting how proteins will fold. While the future of AI development is inherently difficult to predict, we can say that if AI development continues at its current pace, we'll soon face AI powerful enough to fundamentally transform society.

In response, our AI Clarity team at Convergence Analysis recently launched a project focused on analyzing and strategizing for scenarios in which such transformative AI

emerges within the next few years. We believe this threshold of transformative AI (TAI) is a very useful milestone for exploring AI scenarios. This is due to TAI's:

- Wide but well-defined scope;
- Focus on societal impacts of advanced AI;
- Lack of dependence on any specific form of AI.

We don't know what form TAI systems will take, but we can analyze which scenarios are most likely and which strategies are most effective across them. This analysis is important for AI safety. The timelines to TAI should determine our research priorities: if TAI is a century away, we have a lot of time to research and prepare for it. If it's two years away, we may need immediate and drastic global action to prevent calamity.

What is TAI?

We'll start by consulting the existing AI literature to explore various definitions of TAI.

Gruetzemacher and Whittlestone (2019) identify TAI as a useful milestone for discussing AI policy, pointing out that "the notion of transformative AI (TAI) has begun to receive traction among some scholars (Karnofsky 2016; Dafoe 2018)". They argue that the term reflects "the possibility that advanced AI systems could have very large impacts on society without reaching human-level cognitive abilities." They do point out that the term is, or at least was, under-specified: "To be most useful, however, more analysis of what it means for AI to be 'transformative' is needed". In particular, they define TAI as "Any AI technology or application with potential to lead to practically irreversible change that is broad enough to impact most important aspects of life and society".

Similarly, Karnofsky (2021) defines TAI as "AI powerful enough to bring us into a new, qualitatively different future, and in The Direct Approach framework developed by the Epoch team, they define TAI as "AI that if deployed widely, would precipitate a change comparable to the industrial revolution".

Maas (2023) surveys the literature more broadly, exploring various approaches for defining advanced AI. Maas identifies four general approaches: 1) form and architecture of advanced AI, 2) pathways towards advanced AI, 3) general societal impacts of advanced AI, and 4) critical capabilities of particular advanced AI systems. For our scenario analysis, we are interested in both the pathways towards advanced AI (which help us

understand key variables for scenario planning) and the general societal impacts of advanced AI. Of course, the form of advanced AI and its capabilities will also be relevant to our overall analysis. You can explore Maas' overview for a comprehensive high-level take on these general approaches.

Within the approach of understanding the general societal impact of advanced AI, Maas identifies uses of the term "TAI" in influential reports across AI safety and AI Governance. For example, he finds that TAI is the favored term in recent reports from Open Philanthropy and Epoch, as noted above. Maas describes TAI as a definition based on socio-technical change, focusing more on the societal impacts of advanced AI and less on the specific architecture of AI or philosophical questions around AI.

Maas identifies selected themes and patterns in defining TAI. These include:

- Significant, irreversible changes broad enough to impact all of society; possibly precipitates a qualitatively different future
- Transition comparable with the agricultural or industrial revolutions

Building on Maas' themes here, and using the agricultural and industrial revolutions as our loose benchmarks for what would be considered transformative, we define TAI in the following way:

Transformative AI (TAI) is AI that causes significant, irreversible changes broad enough to impact all of society.

This threshold of TAI is related to, but distinct from, other thresholds like artificial superintelligence (commonly ASI) or other levels of Artificial General Intelligence, AGI. These thresholds generally refer to *capabilities* of specific AI systems that surpass human capability in most domains, if not all. However, AI could still transform society without reaching those specific milestones. Here are a few examples of transformative AI scenarios that do not require a high level AGI (say level 4 or level 5 as DeepMind defines it) or ASI:

- AI automates a large fraction of current tasks, leading to mass unemployment and possible economic chaos.
- Narrow AI revolutionizes energy production and distribution, resulting in an end to scarcity and poverty.
- A totalitarian state uses AI to defeat rivals and maintain an extremely stable dystopian regime.

- A malicious actor uses advanced AI to develop and distribute an incredibly virulent virus, killing nearly all humans.

These examples don't require general or super intelligence, but they're still revolutionary. In summary, AI can be societally transformative without crossing the thresholds of AGI or ASI..

Revolutions, Competency, and the Automation Benchmark

Above, we've focused on various definitions of TAI from the literature. However, there are several related concepts that help further illustrate what TAI is and how we might know when it has arrived. In this section, we'll explore Garfinkel's notion of revolutionary technologies, a threshold for TAI identified by Deepmind, and a brief discussion on the most favored benchmark for when TAI will have arrived.

Revolutionary Technology

What constitutes a *revolutionary technology*? Garfinkel's The Impact of Artificial Intelligence: A Historical Perspective (2022) considers "revolutionary technologies" to be a subset of the broader category of "general purpose technologies". For Garfinkel, general-purpose technologies are "distinguished by their unusually pervasive use, their tendency to spawn complementary innovations, and their large inherent potential for technical improvement", while revolutionary technologies are general purpose technologies that "[support] an especially fundamental transformation in the nature of economic production." Garfinkel lists domesticated crops and the steam engine as two particularly noteworthy revolutionary technologies helping to trigger the neolithic and industrial revolutions respectively.

Garfinkel argues that AI may be revolutionary through task automation, and suggests "near-complete automation" as a benchmark for when AI would have revolutionized society. Indeed, this threshold is common in surveys of the public and expert opinion. These surveys often frame TAI-arrival questions as something like "When do you think AI will be as good as humans at X% of tasks?", with X ranging between 80 and 100%.

Competent AGI

We like “TAI” due to its focus on socio-technical impact rather than pure capability thresholds, but measuring capability is still relevant to TAI scenario analysis, especially when AI capabilities are measured in comparison to human performance and societal automation. For example, DeepMind recently provided a useful classification of AGI in terms of generality and performance. In particular, they divide AI capability into six levels (and provide narrow and general examples of each, which we’ll omit):

- Level 0: No AI.
- Level 1: Emerging - performing equal to or slightly better than an unskilled person.
- Level 2: Competent - performing better than at least 50% of skilled adults.
- Level 3: Expert - performing better than at least 90% of skilled adults.
- Level 4: Virtuoso - performing better than at least 99% of skilled people.
- Level 5: Superhuman - outperforming everyone.

The DeepMind team write that:

> The "Competent AGI" level, which has not been achieved by any public systems at the time of writing, best corresponds to many prior conceptions of AGI, and may precipitate rapid social change once achieved.

While the DeepMind team keeps their focus on AGI, they seem to believe that "rapid social change" would be precipitated somewhere around the level of "competent AGI". This plausibly corresponds to a threshold for TAI. Interestingly, for the DeepMind team, this threshold is not “near total automation” but rather “performing better than at least 50% of skilled adults.”

The Task Automation Benchmark

In our current context, advanced AI systems are becoming increasingly general, capable, and autonomous in ways that are already changing the tasks that are of value for humans to complete. In the past, this change in the nature of task completion precipitated transformative change. It seems natural that TAI will bring about its transformation through the same general mechanism of task automation.

While neither Garfinkel nor the DeepMind team are explicitly using the term TAI, they are both pointing at AI systems that would be revolutionary and precipitate rapid social

change. To assess the presence of this sort of AI system, they both suggest (as did we earlier in this post) task automation as a measuring stick. For Garfinkel this is "near total automation" and for DeepMind it's "performing better than at 50% of skilled adults" which implies significant task automation. For our part, we suggest that TAI could transform society when "AI automates a large fraction of current tasks."

Note that the agricultural and industrial revolutions - our guiding historic examples of societal transformation - were both also precipitated by the use of new technology to change the tasks that are of value to complete in the economy. In the agricultural revolution humans began to make use of simple farming tools, often with the aid of energy provided by "beasts of burden." The industrial revolution shifted a major source of power and energy away from humans and various land animals to steam and electric powered machines, which mechanically automated swaths of economically productive tasks and generated countless new ones as well.

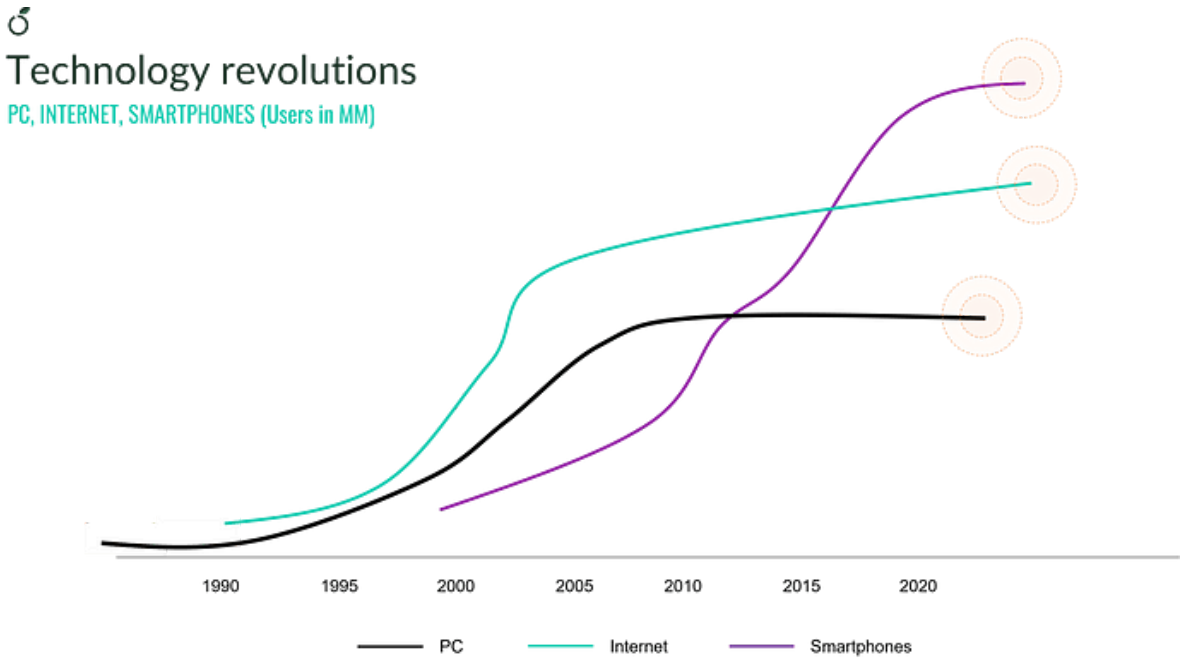
The date that TAI is developed is not the date TAI transforms society, but it may happen fast

The date of development of TAI is not the date TAI transforms society, just as "The date of AI Takeover is not the day AI takes over". AI frontier systems are created through a research and development process. With LLMs, this process includes things like gathering, cleaning, and preparing data for training, conducting the training run, fine tuning, narrow beta release, feedback, wide release, then iterating. Frontier systems then serve as a core component to a wide range of applications. These applications are then deployed and used by organizations and individuals, and *with time*, they change how tasks are performed.

For AI scenario planning, this *with time* part is an important component for understanding which particular scenarios may be most likely. In the next post in this series, our colleague Zershaaneh Qureshi will explore the literature on the date of arrival of TAI. This literature is, generally speaking, exploring when TAI will be developed. However, the date of development of TAI is not the date TAI transforms society. But, it may happen fast.

It may happen fast because the speed of technological adoption is increasing.

In the past, it has often taken significant time for a new technology to: 1) be widely adopted and 2) become the norm for widespread task completion. For example, here's a classic S-shaped curve for PC, Internet, and Smartphone adoption from Medium:



However, the timeline from when a new AI technology is first developed, then deployed, and then adopted for widespread significant use may be measured in months to a couple of years. For example, "ChatGPT, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the fastest-growing consumer application in history." 100 million users remains a small percentage of people in a world of 8 billion humans, but the adoption line is steep.

It may happen fast because TAI comes in the form of AGI agent(s)

As we have emphasized throughout, TAI may arise without the development of AGI or AGI agents. It may be the case that TAI is achieved through very capable narrow AI systems or various sets of comprehensive AI services. In these cases we might expect fast adoption as with ChatGPT, but with TAI in the form of powerful AGI agent(s) "adoption" is the wrong lens. If TAI comes in the form of AGI agent(s) then the pace of transformation is likely to be a function of the AGI agent's motivations and capabilities. What does the AGI want°? What capabilities does it have? Is it capable of recursive self improvement and resource acquisition°?

The *with time* part does matter, but time may be short in the age of TAI.

Looking ahead

In our team's next two posts, we tackle these topics in more detail. First, Zershaaneh Qureshi will provide a detailed overview of *timelines to TAI*. This post will primarily explore the question: when could TAI emerge? Following this, Corin Katzke will explore the topic of AI *agency*. That is, should we expect TAI to be agentic°? This will shed more light on whether we should expect TAI to actually transform society by its widespread adoption or by TAI acting willfully and effectively to achieve its own goals, bending the shape of the transformation to its will.

Convergence Analysis (org) 1 | Transformative AI 1 | AI 2 | Frontpage

Mentioned in

57 Now THIS is forecasting: understanding Epoch's Direct Approach

16 AI Clarity: An Initial Research Agenda

You cannot comment at this time (Questions? Send an email to team@lesswrong.com)

Moderation Log