BROOKINGS

Report

Why we need a new agency to regulate advanced artificial intelligence: Lessons on AI control from the Facebook Files

Anton Korinek Wednesday, December 8, 2021

With the development of ever more advanced artificial intelligence (AI) systems, some of the world's leading scientists, AI engineers and businesspeople have expressed concerns that humanity may lose control over its creations, giving rise to what has come to be called the AI Control Problem. The underlying premise is that our human intelligence may be outmatched by artificial intelligence at some point and that we may not be able to maintain meaningful control over them. If we fail to do so, they may act contrary to human interests, with consequences that become increasingly severe as the sophistication of AI systems rises. Indeed, recent revelations in the so-called "Facebook Files" provide a range of examples of one of the most advanced AI systems on our planet acting in opposition to our society's interests.

In this article, I lay out what we can learn about the AI Control Problem using the lessons learned from the Facebook Files. I observe that the challenges we are facing can be distinguished into two categories: the technical problem of *direct control* of AI, i.e. of ensuring that an advanced AI system does what the company operating it wants it to do, and the governance problem of *social control* of AI, i.e. of ensuring that the objectives that companies program into advanced AI systems are consistent with society's objectives. I analyze the scope for our existing regulatory system to address the problem of social control in the context of Facebook but observe that it suffers from two shortcomings. First, it leaves regulatory gaps; second, it focuses excessively on after-the-fact solutions. To pursue a broader and more pre-emptive approach, I argue the case for a new regulatory body—an AI Control Council—that has the power to both dedicate resources to conduct research on the direct AI control problem and to address the social AI control problem by proactively overseeing, auditing, and regulating advanced AI systems.

What is the AI control problem?

A fundamental insight from control theory[1] is that if you are not careful about specifying your objectives in their full breadth, you risk generating unintended side effects. For example, if you optimize just on a single objective, it comes at the expense of all the other objectives that you may care about. The general principle has been known for eons. It is reflected for example in the legend of King Midas, who was granted a wish by a Greek god and, in his greed, specified a single objective: that everything he touched turn into gold. He realized too late that he had failed to specify the objectives that he cared about in their full breadth when his food and his daughter turned into gold upon his touch.

The same principle applies to advanced AI systems that pursue the objectives that we program into them. And as we let our AI systems determine a growing range of decisions and actions and as they become more and more effective at optimizing their objectives, the risk and magnitude of potential side effects grow.

The revelations from the Facebook Files are a case in point: Facebook, which recently changed its name to Meta, operates two of the world's largest social networks, the eponymous Facebook as well as Instagram. The company employs an advanced AI system —a Deep Learning Recommendation Model (DLRM)—to decide which posts to present in the news feeds of Facebook and Instagram. This recommendation model aims to predict which posts a user is most likely to engage with, based on thousands of data points that the company has collected about each of its billions of individual users and trillions of posts.

Facebook's AI system is very effective in maximizing user engagement, but at the expense of other objectives that our society values. As revealed by whistleblower Frances Haugen via a series of articles in the *Wall Street Journal* in September 2021, the company repeatedly prioritized user engagement over everything else. For example, according to Haugen, the company knew from internal research that the use of Instagram was associated with serious increases in mental health problems related to body image among female teenagers but did not adequately address them. The company attempted to boost "meaningful social interaction" on its platform in 2018 but instead exacerbated the

promotion of outrage, which contributed to the rise of echo chambers that risk undermining the health of our democracy. Many of the platform's problems are even starker outside of the U.S., where drug cartels and human traffickers employed Facebook to do their business, and Facebook's attempts to thwart them were insufficient. These examples illustrate how detrimental it can be to our society when we program an advanced AI system that affects many different areas of our lives to pursue a single objective at the expense of all others.

The Facebook Files are also instructive for another reason: They demonstrate the growing difficulty of exerting control over advanced AI systems. Facebook's recommendation model is powered by an artificial neural network with some 12 trillion parameters, which currently makes it the largest artificial neural network in the world. The system accomplishes the job of predicting which posts a user is most likely to engage with better than a team of human experts ever could. It therefore joins a growing list of AI systems that can accomplish tasks that were previously reserved for humans at super-human levels. Some researchers refer to such systems as domain-specific, or narrow, superintelligences, i.e. AI systems that outperform humans within a narrow domain of application. Humans still lead when it comes to general intelligence—the ability to solve a wide range of problems in many different domains. However, the club of narrow superintelligences has been growing rapidly in recent years. It includes AlphaGo and AlphaFold, creations of Google subsidiary DeepMind that can play Go and predict how proteins fold at super-human levels, as well as speech recognition and image classification systems that can perform their tasks better than humans. As these systems acquire super-human capabilities, their complexity makes it increasingly difficult for humans to understand how they arrive at solutions. As a result, an AI's creator may lose control of the AI's output.

Direct and social control over our AI systems

There are two dimensions of AI control that are useful to distinguish because they call for different solutions: The direct control problem captures the difficulty of the company or entity operating an AI system to exert sufficient control, i.e. to make sure the system does what the operator wants it to do. The social control problem reflects the difficulty of ensuring that an AI system acts in accordance with social norms.

Direct AI control is a technical challenge that companies operating advanced AI systems face. All the big tech companies have experienced failures of direct control over their AI systems—for example, Amazon employed a resume-screening system that was biased against women; Google developed a photo categorization system that labeled black men as gorillas; Microsoft operated a chatbot that quickly began to post inflammatory and offensive tweets. At Facebook, Mark Zuckerberg launched a campaign to promote COVID-19 vaccines in March 2021, but one of the articles in the Facebook Files documents that Facebook instead turned into a source of rampant misinformation, concluding that "[e]ven when he set a goal, the chief executive couldn't steer the platform as he wanted."

One of the fundamental problems of advanced AI systems is that the underlying algorithms are, at some level, black boxes. Their complexity makes them opaque and makes their workings difficult to fully understand for humans. Although there have been some advances in making deep neural networks explainable, these are innately limited by the architecture of such networks. For example, with sufficient effort, it is possible to explain how one particular decision was made (called local interpretability), but it is impossible to foresee all possible decisions and their implications. This exacerbates the difficulty of controlling what our AI systems do.

Frequently, we only detect AI control problems after they have occurred—as was the case in all the examples from big tech discussed above. However, this is a risky path with potentially catastrophic outcomes. As AI systems acquire greater capabilities and we delegate more decisions to them, relying on after-the-fact course corrections exposes our society to large potential costs. For example, if a social networking site contributes to encouraging riots and deaths, a course correction cannot undo the loss of life. The problem is of even greater relevance in AI systems for military use. This creates an urgent case for proactive work on the direct control problem and public policy measures to support and mandate such work, which I will discuss shortly below.

Social control over AI and governance

In contrast to the technical challenge of the direct control problem, the social AI control problem is a governance challenge. It is about ensuring that AI systems—including those that do precisely what their operators want them to do—are not imposing externalities on the rest of society. Most of the problems identified in the Facebook Files are examples of this, as Zuckerberg seems to have prioritized user engagement—and by extension the profits and market share of his company—over the common good.

The problem of social control of AI systems that are operated by corporations is exacerbated by market forces. It is frequently observed that unfettered market forces may provide corporations with incentives to pursue a singular objective, profit maximization, at the expense of all other objectives that humanity may care about. As we already discussed in the context of AI systems, pursuing a single objective in a multi-faceted world is bound to lead to harmful side effects on some or all members of society. Our society has created a rich set of norms and regulations in which markets are embedded so that we can reap the benefits of market forces while curtailing their downsides.

Advanced AI systems have led to a shift in the balance of power between corporations and society—they have given corporations the ability to pursue single-minded objectives like user engagement in hyper-efficient ways that used to be impossible before such technologies were available. The resulting potential harms for society are therefore larger and call for more proactive and targeted regulatory solutions.

A proposal to establish an AI Control Council

Throughout our history, whenever we developed new technologies that posed new hazards for society, our nation has made it a habit to establish new regulatory bodies and independent agencies endowed with world-class expertise to oversee and investigate the new technologies. For example, the National Transportation Safety Board (NTSB) and the Federal Aviation Administration (FAA) were established at the onset of the age of aviation; or the Nuclear Regulatory Commission (NRC) was established at the onset of the nuclear

age. By many measures, advanced artificial intelligence has the potential to be an even more powerful technology that may impose new types of hazards on society, as exemplified by the Facebook Files.

Given the rise of artificial intelligence, it is now time to establish a federal agency to oversee advanced artificial intelligence—an AI Control Council that is explicitly designed to address the AI Control Problem, i.e. to ensure that the ever more powerful AI systems we are creating act in society's interest. To be effective in meeting this objective, such a council would need to have the ability to (i) pursue solutions to the direct AI control problem and (ii) to oversee and when necessary regulate the way AI is used across the U.S. economy to address the social control problem, all while ensuring that it does not handicap advances in AI. (See also here for a complementary proposal by Ryan Calo for a federal agency to oversee advances in robotics.) In what follows I first propose the role and duties of an AI Control Council and then discuss some of the tradeoffs and design issues inherent in the creation of a new federal agency.

First, there are many difficult technical questions related to direct AI control—and even some philosophical questions—that require significant fundamental research. Such work has broad public benefits but is hampered by the fact that the most powerful computing infrastructure, the most advanced AI systems, and increasingly the vast majority of AI researchers are located within private corporations which do not have sufficient incentive to invest in broader public goods. The AI Control Council should have the ability to direct resources to addressing these questions. Since the U.S. is one of the leading AI superpowers, this would have the potential to steer the direction of AI advancement in a more desirable direction at a worldwide level.

Second, to be truly effective, the council would need to have a range of powers to oversee AI development by private and public actors to meet the challenge of social control of AI:

1. It should have the power to monitor AI development and define which types of advanced AI systems, in the private sector and elsewhere, fall under the regulatory oversight of the Council. It can base this assessment on criteria such as the size of neural networks, the amount of compute employed (i.e. the resources used for computation), the reach of the systems (e.g., how many people they interact with and

how wide-ranging their effects are anticipated to be), or other criteria that the Council deems appropriate.

2. It should have the power to mandate impact assessments of these advanced AI systems on a variety of stakeholders; the ability to define what yardsticks advanced AI companies need to report on; and the capacity to perform audits and experiments to ascertain their impacts in the real world. These impact assessments and the related questions and experiments would need to differ significantly depending on the type of AI systems and the concerns that they raise. For example, a social network may be asked to report on all the areas of concern that have been discussed in the context of the Facebook Files, ranging from content moderation and fairness concerns to its impact on the mental health of its users and on democratic discourse. Tools to supercharge biomedical research, such as AlphaFold, may be asked to evaluate the potential for abuse by creating novel pathogens. Advanced language models such as GPT-3 that can generate large quantities of human-level language may be asked to evaluate their effects on targeted consumer manipulation and misinformation.
3. When the impact assessments indicate risks to society or potential abuses, the Council needs the regulatory powers to curtail these risks and abuses as well as the power to supervise and enforce the implementation of any remedies or regulations that result.
4. The lessons from the impact assessments should be publicly available to increase the transparency of advanced AI systems and to raise awareness of the potential problems to look out for—not only among consumers and workers, but also among other AI developers that may deal with similar problems. Another benefit of transparency is that it helps consumers, workers, and venture capitalists decide which companies to support and which ones to steer clear of if some AI companies prioritize narrow objectives to the detriment of the broader objectives of our society.

Challenges establishing an AI Control Council

Since talent shortages in the AI sector are severe, the Council needs to be designed with an eye towards making it attractive for the world's top experts on AI and AI control to join. Many of the leading experts on AI recognize the high stakes involved in AI control. If the

design of the Council carries the promise to make progress in addressing the AI control problem, highly talented individuals may be eager to serve and contribute to meeting one of the greatest technological challenges of our time.

One of the questions that the Council will need to address is how to ensure that its actions steer advances in AI in a desirable direction without holding back technological progress and U.S. leadership in the field. The Council's work on the direct control problem as well as the lessons learned from impact assessments will benefit AI advancement broadly because they will allow private sector actors to build on the findings of the Council and of other AI researchers. Moreover, if well-designed, even the oversight and regulation required to address the social control problem can in fact spur technological progress by providing certainty about the regulatory environment and by forestalling a race to the bottom by competing companies.

Another important question in designing the Council is resolution of domain issues when AI systems are deployed in areas that are already regulated by an existing agency. In that case, it would be most useful for the Council to play an advisory role and assist with expertise as needed. For example, car accidents produced by autonomous vehicles would fall squarely into the domain of the National Highway Traffic Safety Administration (NHTSA), but the new AI Control Council could assist with its expertise on advanced AI.

By contrast, when an advanced AI system gives rise to (i) effects in a new domain or (ii) emergent effects that cut across domains covered by individual agencies, then it would fall within the powers of the AI Control Council to intervene. For example, the mental health effects of the recommendation models of social networks would be a new domain that is not covered by existing regulations and that calls for impact assessments, transparency, and potentially for regulation. Conversely, if for example a social network targets stockbrokers with downbeat content to affect their mood and by extension stock markets to benefit financially in a way that is not covered by existing regulations on market manipulation, it would be a cross-domain case that the council should investigate alongside the Securities and Exchange Commission (SEC).

Conclusion

From a longer-term perspective, the problems revealed in the Facebook Files are only the beginning of humanity's struggle to control our ever more advanced AI systems. As the amount of computing power available to the leading AI systems and the human and financial resources invested in AI development grow exponentially, the capabilities of AI systems are rising alongside. If we cannot successfully address the AI control problems we face now, how can we hope to do so in the future when the powers of our AI systems have advanced by another order of magnitude? Creating the right institutions to address the AI control problem is therefore one of the most urgent challenges of our time. We need a carefully crafted federal AI Control Council to meet the challenge.

The Brookings Institution is financed through the support of a diverse array of foundations, corporations, governments, individuals, as well as an endowment. A list of donors can be found in our annual reports published online here. The findings, interpretations, and conclusions in this report are solely those of its author(s) and are not influenced by any donation.

Footnotes

1. 1 Control theory is a branch of engineering that describes how to develop systems that pursue desired outcomes.