AGI and Democracy

Seth Lazar and Alex Pascal

MARCH 2023

ALLEN LAB FOR DEMOCRACY RENOVATION

HARVARD Kennedy School
ASH CENTER
for Democratic Governance
and Innovation

2023 was the year of AI, as new products brought significant recent advances in the field to universal attention, and the world's most powerful tech companies declared their ambitions to achieve Artificial General Intelligence (AGI). 2024 will be the year of democratic elections, with a record-breaking 40-plus countries (including the U.S., India, U.K., Ukraine, Taiwan, and South Africa), representing more than 40% of the world's population, going to the polls. Already, many are justifiably fretting about the direct impacts of AI on democracy, as our information and communication environment becomes ever more polluted with AI-generated deepfakes, disinformation, and potential cyberattacks. How much of a difference AI will make in 2024 remains an open question. However, beyond these immediate threats the advent of AGI could finish democracy once and for all.

To many, this fear will seem remote, perhaps hypothetical. That's understandable. There are, at present, many more pressing concerns raised by existing AI systems. These include a range of harms, such as bias and discrimination, worker exploitation, turbo-charged online abuse and scamming, mass surveillance, and the further erosion of privacy. On the horizon is the potential for significant economic disruption from labor displacement, and the collapse of the creative economy (and potentially the internet as we know it). Most of AI's problems today stem from how it falls short of expectations, not from how astonishingly powerful it is.

So, if we are a long way short of AGI, why worry about it now? Because the people building the most advanced AI systems are explicitly and aggressively working to bring AGI about, and they think they'll get there in two to five years. Even some of the most publicly skeptical AI researchers don't rule out AGI within this decade. If we, the affected public, do not actively shape this agenda now, we may miss the chance to do so at all. We face a fundamental question: is the *very pursuit* of AGI the kind of aim democracies should allow?

Defining AGI

But what even is AGI? Defining it sometimes feels like pinning Jell-O to a wall. But as progress accelerates, something like a consensus is emerging. Synthesising a vast literature, we can say that AGI would be a nonbiological computational system that can perform any cognitive function currently performed by humans at the level of the median human or better (acknowledging the crude quantification this implies). Google DeepMind's recent paper mentions "linguistic intelligence, mathematical and logical reasoning, spatial reasoning, interpersonal and intra-personal social intelligences, the ability to learn new skills and creativity." Other AI researchers would also add instrumental rationality, causal reasoning, tool use, and at least some ability to distinguish truth from falsehood. OpenAI calls it, simply, "AI systems that are generally smarter than humans."

Existing AI systems are undoubtedly far from AGI across all these criteria, besides perhaps linguistic competence. And yet, GPT-4, OpenAI's most advanced model, is significantly more general and capable than earlier systems, and other AI companies have in the last year caught up and (barely) surpassed GPT-4 in many respects. The feasibility horizon of AI research is rapidly expanding outward. And while we do not yet have AGI's ingredients, many think we know where to look—in terms of both building on GPT-4's successes and backfilling its limitations. What's more, the leading AI labs and Big Tech companies—including five of the world's seven most valuable companies—have an explicit mission to achieve AGI. Whether you think that's just hype or else that the advent of AGI is inevitable, we should at least ask, *now*, whether pursuing that goal is itself consistent with democratic values.

A Democratic Greenlight, Not Just Guardrails

At a first pass, the answer seems to be no. AGI could do more than any preceding innovation to shape and disrupt our economies, politics, culture, and communities. In democracies, the people are sovereign. All should stand as equals and govern together (at least through our representatives). There is nothing inevitable about AGI's arrival. It is a choice. One that will affect all of us profoundly. The question is, who's making it? Right now, the answer is a few people at very few companies. Allowing these companies to unilaterally pursue the development of technologies as potentially transformative as AGI is self-evidently undemocratic.

Despite the well-intentioned experiments in corporate governance to make some leading AI labs more public-interested than regular businesses, a board of directors simply cannot adequately represent the societies and people whose trajectories and lives AGI will radically transform. Thus far, the public debate about how to rein in the societal impacts of AI has focused only on identifying guardrails to shape its development, mainly to mitigate harms and risks. This is necessary but not sufficient. We also need to ask the more fundamental question of whether we actually want to build AGI in the first place. The advent of future technologies is not a foregone; for example, with human cloning, we have shown that we can slow or stop development if we choose to do so. Whatever you think our AI future should be, it should be one that we have consciously chosen together.

A Democratic Path to AGI?

Suppose then that democratic publics explicitly and affirmatively decide that they want AGI. Could we develop it in accordance with democratic values? Some of the leading AI labs clearly recognize this question's urgency and are making respectable efforts in this direction. Incorporating democratic inputs into AI development has already led to some noteworthy improvements (and the authors welcome attempts to use AI more generally to enhance political participation). But democratic inputs are not the same as democratic control. Accepting inputs presupposes controlling the agenda and dictating where inputs are welcome. There will be many branching paths on the road to AGI, and at many of those junctures, such as determining how to source and filter the data used to train AI systems, the public interest will predictably conflict with the pursuit of returns on investment. When tens of billions of dollars have been invested in a company, those billions will ultimately set the agenda, irrespective of corporate structure. OpenAI's recent governance crisis is case in point.

In addition to the kind of deep democratic oversight of AI development that some have proposed, the only reliably democratic path to AGI would likely involve complementing any private sector research and development with a robust and capable counterpart driven purely by the public interest—an AI "public option," as some have called it. Investments like the National AI Research Resource could light a path to such an approach but would require an order of magnitude greater commitment and ambition to succeed.

What Happens If We "Succeed"?

But does the path to AGI lead somewhere that democracies should go? In 2023, many loudly argued "no," not because of the implications for democracy, but because they considered AGI an existential threat to humanity at large. People have presented scenarios, ranging from speculative to compelling, in which AGI is humanity's final, civilization-ending invention. The frontier labs also feel this critique keenly and have built research teams aiming to "align" AGI (and ASI, artificial superintelligence, its successor) to make it beneficial, safe and controllable. But even if we can align AGI to mitigate existential

risk (assuming we can democratically decide what that means and how to do it), it would still not be enough for AGI to be deemed safe for democracy.

Here's why: If AGI is better than most humans at all cognitive tasks, it is very likely to be better than humans at the numerous tasks of governing—that is, designing, implementing, and enforcing the rules by which a community or institution operates. This will create a compelling incentive to invest AGI with governing power at all levels of society, from clubs, schools, and workplaces to the administrative agencies that regulate and help steward the economy, labor, the environment, transport, health care, and even provide for public safety, criminal justice, and election administration. If in fact AGI is much better at executing the tasks that we give it than humans (as its would-be creators intend), there will be a strong, perhaps irresistible temptation to have it identify and select which tasks to pursue, then to have it set our priorities, not just make and enforce our rules in particular domains.

As new threats and problems arise faster than we can process them, we may very well entrust AGI with a blanket authority to prioritize, decide and act on our behalf. We would de facto be kissing good-bye to democracy in any real sense of its value and practice. Think of this threat as an absent-minded walk down a political primrose path, not the more widely-discussed 'rogue AI' scenarios. Do we want a world in which we abdicate to AI our power and responsibility to determine the values and priorities that guide our societies because of its sheer performance capacity? We already see this kind of easy deference to our existing, deeply flawed computational systems. It would only be exacerbated, perhaps irreversibly so, with AGI.

From decades of work on automation, we know that in every domain, from manufacturing to algorithmic trading, automating a task and then relying on humans for oversight at critical moments is a doomed project. The goal of making future AGI systems "controllable" cannot be achieved through technology design alone. For *anything* to be controllable, we must presuppose that something or someone is doing the controlling. It is not enough to design systems that could, in principle, be controlled, when we can reliably predict, based on past experience, that humans will fail to use the controls that we have designed for them. Nor is having some AGIs control others an adequate answer. For AGI to be safe for democracy, *democratic institutions* run by people must be able and expected to exercise meaningful control. This may well require rethinking the aging institutions of constitutional democracy itself—something that only we the People can legitimately do.

Where Next?

Setting AI entirely aside, this year will prove for many democracies their sternest test yet and may see more voters than ever before choose candidates who have explicitly promised an anti-democratic agenda. These developments show that we cannot take the value of democracy for granted, treating it as such a sacrosanct and shared ideal that nobody could ever credibly make an argument against it. Some might embrace the idea of replacing our messy, disputatious political systems with "efficient," "impartial," "optimizing" technocratic AGI rule. Many others—the authors included—will not. But let's have that debate. And let's not underestimate the gravity of the choice we're already making passively, by default. Otherwise, in 2024 we might save democracy from the would-be autocrats, only to pave the way for AGI to deliver it an even more lethal blow.

About the Author

Seth Lazar is Professor of Philosophy at the Australian National University, and a Distinguished Research Fellow of the University of Oxford Institute for Ethics in AI. He writes about the moral and political philosophy of computing; his *Connected by Code: How AI Structures, and Governs, the Ways We Relate*, based on his 2023 Tanner Lecture on AI and Human Values, is forthcoming with Oxford University Press. His recent work can be found at linktr.ee/sethlazar.

Alex Pascal is a Senior Fellow at the Ash Center for Democratic Governance and Innovation at the Harvard Kennedy School of Government. As Special Assistant to the President for Domestic Policy from January 2021 through June 2023, Alex helped lead Biden Administration policy initiatives on both artificial intelligence and democracy.

About the Ash Center

The Mission of the Roy and Lila Ash Center for Democratic Governance and Innovation at is to develop ideas and foster practices for equal and inclusive, multi-racial and multi-ethnic democracy and self-government.

This essay is one in a series published by the Ash Center for Democratic Governance and Innovation at Harvard University's John F. Kennedy School of Government. The views expressed in this essay are those of the authors and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. The papers in this series are intended to elicit feedback and to encourage debate on important public policy challenges.

A PUBLICATION OF THE

Ash Center for Democratic Governance and Innovation
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138
617-495-0557
ash.harvard.edu

A PUBLICATION OF THE

Ash Center for Democratic Governance and Innovation
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

617-495-0557
ash.harvard.edu

HARVARD Kennedy School
ASH CENTER
for Democratic Governance
and Innovation