KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

AI Agents and Democratic Resilience

Seth Lazar and Mariano-Florentino Cuéllar

September 2025

Sébastien A. Krier using Midjourney 6.1

Introduction

Computational progress has always been Janus-faced for democracy. The spread and networking of computing power bolster the epistemic and communicative practices on which democracies rely [30, 38, 164]. Yet the same tools are among the most sophisticated instruments of coercion and control ever devised [47, 78, 86, 166].

Every landmark in computing—from the first digital machines to the PC, the internet, and now artificial intelligence (AI)—has provoked anguished reassessment of this tension [1, 16, 21, 29, 93, 99, 101, 120, 132, 157, 159, 160]. In 1984, Langdon Winner [159] described an enduring divide: "computer romantics" who dream that each leap forward will finally realize computing's unkept promise for democracy; and skeptics, like himself, who think that more powerful technologies always serve the powerful first, best, and perhaps only.

For computing's first half-century, the romantics seemed to have the better of the argument. Computational and democratic progress proceeded hand in hand. In the 21st century,

however, this picture has darkened. Democratic ideals face acute pressure: just over a quarter of humanity now lives in electoral or liberal democracies, down from almost half in 2016.[1] Countries sliding toward autocracy now double those moving toward democracy.[2] And while computational progress has accelerated, the public's endorsement of the social role of computing in general, and of technology companies in particular, has recently faltered [98]. Through the mid-2010s, big-tech companies ranked among society's most trusted institutions [36]. Since then, a cross-national backlash against platform power and digital harms [53] has spurred heavy regulation and, even in the United States, a bipartisan conviction that too few companies wield too much power.[3]

Policymakers and the public today face another digital revolution. In the last decade, research progress in AI has taken off [17, 84, 104, 149]. We have already developed extraordinarily powerful and economically valuable analytical and generative AI tools. We are now on the cusp of building autonomous AI systems that can carry out almost any task that competent humans can currently use digital technologies to perform. Our democracies will soon be infused with AI agents.

In this paper, we explore how AI agents might benefit, advance, and complicate the realization of democratic values. We aim to consider both faces of the computational Janus, avoiding both Panglossian optimism and ahistorical catastrophizing.

We begin (Section 2) by defining key terms and introducing our approach. We then explore AI agents' democratic implications through three lenses. Section 3 examines how agents may interact with structural pressures already straining democratic institutions. Section 4 identifies novel threats they could introduce. Section 5 outlines how to design 'agents for democracy' that reinforce, rather than undermine, those institutions.

Democracies are weaker than they have been for decades. A great wave is coming, and they are ill-prepared. AI agents may be cure as well as cause, but we cannot depend on them, nor can we simply trust that they will advance democratic values by default. Our urgent task is to rebuild and revitalize the institutions and practices that advance democratic values, restoring their resilience against the technological and social upheaval ahead [13].

Groundwork

Language Model Agents

AI agents are computational systems that can independently pursue relatively long and complex sequences of actions towards a goal, through functionally understanding their

1 In 2024, 2.3 billion people lived in democracies (from a global population of 8.2b). In 2016 the numbers were 3.9b in democracies, 7.5b total [54].

2 In 2024 45 countries were autocratising, 19 democratising, according to Varieties of Democracy and Episodes of Regime Transformation data [54].

3 See, e.g., the complementary antimonopoly visions of the last administration's FTC chair, Lina Khan, and the "conservative antitrust" approach of the current FTC's leading antitrust commissioner, Mark Meador (https://www.ftc.gov/system/files/ftc_gov/pdf/antitrust-policy-for-the-conservative-meador.pdf).

Electronic copy available at: knightcolumbia.org

environment and how it changes when they act, in ways that allow them to dynamically update their plans in response to new information [121].[4] Until recently, the most promising path towards building AI agents seemed to be self-contained reinforcement learning (RL) agents that learn how to operate within an environment through vast amounts of simulation [138]. While this approach proved more successful than some preceding alternatives, one of its key limitations is overreliance on researchers' ability to create a simulated environment precisely matching what the agent will experience once it is deployed [4]. As a result, RL agents are often very brittle and perform poorly in settings that take them out of the distribution on which they were trained [28].[5]

Large language models (LLMs; we include in this category multimodal models based on the same architecture) have offered a new paradigm for AI agent design.[6] Language model agents (LMAs)[7] are compound AI systems in which the LLM functions as the 'executive center' responsible for the control flow of the system, determining how it makes decisions, draws on different modules or subprocesses, repeats actions, or branches into different paths based on conditions.[8]

4 There is a live philosophical (and legal) debate about whether the term 'agent' is warranted here. We intend it in the prosaic sense in which it appears in computer science textbooks, denoting an entity that acts in the pursuit of some goal within an environment that it perceives and from which it receives feedback when it acts. In other work, we consider deeper questions of agency. Thanks to Aziz Huq for pressing us on this point.

5 Obviously, these claims are contested; plenty of computer scientists still think that 'reward is enough' and that RL agents can solve every problem. See, e.g., [126].

6 There is no ideal nomenclature here. For our purposes, 'LLMs' include all large neural networks pretrained on internet-scale data on sequences of tokens to predict masked tokens using self-supervised learning, and then typically post-trained using methods like supervised fine-tuning, RL from human feedback, direct preference optimization, and RL with verifiable rewards. The name 'large *language* model' is, therefore, something of a misnomer. Almost any kind of data can be represented as a sequence of tokens. LLMs, therefore, include multimodal models that natively process audio and image as well as text. The other term with some currency is, of course, 'foundation model,' but this was originally intended to refer to the pretrained model only, emphasizing that it could be used as a foundation for fine-tuning relative to some specific task. This is narrower than we mean to go. The term 'language model agent' inherits these imperfections. However, the kind of AI agents most likely to affect democracies in the near to mid-term will be powered by LLMs, so LMA is an apt name for them. We also note that the linguistic capacity of LMAs is likely a necessary condition for their impacting democracy as we expect them to do because, at least in formal terms, civic life plays out in a linguistic substrate: legislation is written, candidates campaign with words, and political communities describe their ties using language (even if they originate from bonds and reactions that transcend words or even logic).

7 LMAs are sometimes just called language agents but that implies language agents are defined by reasoning *in language*, whereas in fact, LLMs, and a fortiori LMAs, *need not* reason only in language but can instead reason in "continuous latent space" [52].

8 To count as a *language model* agent, the LLM must be involved at least in planning and action selection [134]. Meta's "Cicero," for example, which used LLMs only for communication as an input and output tool, is not an LMA on this definition, because it does not make use of the practical intelligence of LLMs, only their communicative capacities [8].

Electronic copy available at: knightcolumbia.org

These systems make extensive use of reinforcement learning too, but by leveraging the latent practical intelligence of the LLM (its ability to understand and respond to practical reasons), they are much more adaptable than the simple RL agents that they have mostly replaced [153]. Still an LLM on its own is not an agent. It is stateless (lacks memory), its inputs and outputs are narrowly constrained, it cannot initiate action; it can do nothing more than respond to a prompt with some output tokens. To exploit the latent practical intelligence of LLMs, we need to rely on tools and scaffolding [124]. This scaffolding can be divided in different ways, but it plays basically three distinct roles [134].

Perceptual scaffolding is necessary for the LMA to receive information about its environment (including instructions from its principal—the individual or group on whose behalf the agent is acting).[9] This includes their user interface and other techniques for gathering text, audio, and video inputs, as well as any other kinds of data. These are effectively the model's *sensors*.

Their *action* scaffolding is their means for acting on the world. This includes outputting tokens to write code or function calls, which are then run in a code interpreter or an API for some other piece of software. These are the model's *actuators*.

An LMA's *cognitive* scaffolding is the other software that the LLM can use to enhance its practical intelligence. This can include many possibilities, but typical examples are working- and long-term memory, learning, reasoning and planning modules, and verifiers that check code, plans, reasoning, or conformity to safety or ethical principles [63].

Given the role of agents in virtually any scenario involving societal benefits and long-term productivity gains through AI, LMAs are now the focus of the frontier AI labs and a vast ecosystem of start-ups. Until recently, they were not highly performant [65]. For example, on one benchmark designed to test software engineering capability, start-up Cognition was fêted in mid-2024 for reaching a modest 15% with its Devin agent.[10] But things have been changing fast—in late 2024, OpenAI's o3 model achieved 85% on the same test [105]. The most important innovation driving progress is the recognition that reinforcement learning can be used to train LLMs to make better use of "test-time compute"—that is, tokens produced in a chain of thought that provides resources for their final answer [37] (imagine, with a dash of anthropomorphism, that when prompted the model first takes time to write down its thoughts on a scratchpad, allowing for different approaches to be pursued before it presents you with its final response).[11] This has been roughly analogized to investing LLMs with a kind of System II thinking, which allows for better reasoning and planning [60].[12] Besides better performance on benchmarks, this progress in using test-time compute has enabled the development of actual agents that can be deployed on valuable real-world tasks. For example, all of the leading AI companies are now developing or have recently

9 See figure 2.13 in [121].

10 https://www.cognition.ai/blog/swe-bench-technical-report. Devin too has come on in leaps and bounds since it was first introduced.

11 O3 also uses some degree of parallelization and selection over outputs to achieve even higher performance [105].

12 Note, however, some skepticism from [148].

Electronic copy available at: knightcolumbia.org

released LMAs that can use a virtual computer or browser to perform arbitrary tasks at the user's behest (with varying degrees of success [107, 133]). Each leading company also has a "Deep Research" agent, which uses browsers and other tools to conduct extensive research on a given topic [103].

There remains a "capability-reliability" gap for LMAs [65]. And for good theoretical reasons, this gap might persist, so that deploying really consequential agents at scale eludes AI companies for years, just as autonomous driving took years to go from proof of concept to a reliable part of the transportation ecosystem [100]. However, the extraordinary pace of change over the last two years could instead continue, so that people can soon access software agents that are ultimately able to do anything with a computer that a competent human can do—albeit far faster than humans, and massively in parallel.

The range of possible outcomes is overwhelming; to better sort through them, it helps to think about dimensions of AI progress. In [117], Morris et al. focus on *performance* and *generality*. They rank performance intuitively from emerging to competent, expert, virtuoso, and superhuman. Though they divide generality into a binary—narrow and general—this can also be represented as a spectrum. A further dimension is also useful to consider in this context, even if it involves an often-contested concept: *autonomy*, which further divides into *task-autonomy* and *role-autonomy*.[13] The former is the ability to operate without direct human instruction or oversight when completing a particular task. Role-autonomy is the ability to autonomously perform an entire human-equivalent role, which includes, for example, task selection as well as task performance. The resulting distinction obviously exists on a continuum rather than serving as a sharp divide.

These basic resources allow us to specify the target systems whose impacts this paper anticipates. Our analysis is conditioned on the possibility that in the near future companies will develop AI agents that can reach competent to expert human performance in a wide range of tasks (whether on their own or by orchestrating a number of more narrowly capable agents), with substantial task-autonomy but limited role-autonomy. That is, agents that are extremely effective tools, but which still need to be actuated or overseen by at least some humans rather than being able to operate with as much role-autonomy as a human worker. Of course, some human roles require very little role-autonomy, and by automating tasks, agents will make many roles redundant. However, there is an important (albeit blurry) distinction between an agent like Deep Research that is fundamentally a tool, and an agent that is able to spontaneously and independently assign itself tasks to complete.[14]

Anticipatory Ethics

This paper is an exercise in "anticipatory ethics"[62]: the project of identifying ex ante the likely ethical questions raised by new technologies, and using that knowledge to shape those technologies for the better [75].[15] This paper does *not* aim to forecast the net impact

13 https://futureoflife.org/standards

14 Notice that it is possible either to hand off tasks entirely to an agent-as-tool or to work with it collaboratively (our argument is not conditional only on human replacement scenarios).

15 For a complementary discussion of "sociotechnical speculative ethics," see [44]. For a general articulation and defense of anticipatory AI ethics, see [74].

Electronic copy available at: knightcolumbia.org

of LMAs on society, taking into account all of the possible positive and negative effects, weighting them for their probability, and summing them all together. Instead, it aims to identify the features of capable AI agents that, given the environment into which such systems will be deployed, are likely to be either societally beneficial or else harmful. The goal is not to make an all-things-considered prediction, but to highlight discrete hazards and opportunities that can be mitigated or exploited when designing and deploying these systems.

Unlike some approaches to anticipating societal risks from advanced AI, this method is epistemically humble in two respects. First, it confines its inquiry to specific, causally relevant features of the target system, rather than attempting an aggregate judgment of the societal impacts of the systems as a whole. Second, it limits its conclusions to narrow probabilistic claims—how those features raise or lower the likelihood of particular social benefits or harms—without pretending to forecast their eventual net balance. It therefore diverges from (a) forecasting approaches, which aim to assign probabilities to specific societal outcomes (see, for example, prediction markets); (b) 'all-things-considered' approaches that aim to tot up overall social impact (often the path taken by those writing for a wider audience); and (c) scenario methods that spin coherent narrative futures around a presumed median path through the possibility space (a recent trend in the literature on transformative AI, and especially existential risk).[16]

Admittedly, anticipatory ethics is often plagued by a mixture of saliency bias and hyperbolical technological determinism [75]. That is, researchers considering a new technology's prospective societal impacts often focus too narrowly on the technology itself and infer societal impacts from its properties without integrating their analysis into a broader account of existing and causally overlapping trends. This ultimately leads to an exaggerated sense of the importance of that technology with respect to the outcomes in question. For example, in the early days of generative AI, many argued that the ability to generate synthetic content would bring down democracies (e.g., [71]). As it has turned out, democratic institutions are under threat from so many other directions that synthetic content has proved more or less irrelevant so far [127]. This paper aims to avoid this vice by situating our analysis of the democratic impacts of LMAs in a broader account of the trends that are affecting democracies worldwide.

Similarly, if anticipatory ethics becomes unmoored from the details of the specific technologies whose impacts it aims to anticipate and steer, it risks shading into ungrounded speculation where almost anything is possible. The following analysis is therefore intentionally focused on the societal impacts of task-autonomous AI agents, as described above, and neither looks beyond nor short of that mark. Though much of the analysis would carry over to virtuoso role-autonomous agents, as well as to advanced non-agentic AI systems, those fall outside this paper's remit. The same is true for the possibility that AI agents will become superhuman researchers that radically and rapidly advance scientific research [3, 45, 91].[17]

16 To be clear, we make no claim to epistemic *superiority* for our approach to anticipatory ethics over these. One could view greater epistemic humility as a vice, not a virtue.

17 Other work has explored the broader question of AI's impacts on democracy, and in particular, advanced AI's impacts—we have aimed to cover complementary ground in this paper [29, 135].

Electronic copy available at: knightcolumbia.org

'Democracy'

What precisely does it mean for LMAs to impact 'democracy'? On one extremely institutionalist view, LMAs impact democracy if and only if they materially contribute to some democracies becoming non-democracies and vice versa [18, 87, 111]. For example, some political scientists argue against the widespread view that democracy is 'in crisis' on the grounds that there are more democracies today than ever before [144]. As long as a nation-state has pluralistic elections and the peaceful transfer of power, democracy is in good shape—everything else amounts to just disputes *within* democracies, rather than crises *of* democracy [111]. By contrast, we argue that LMAs don't have to materially contribute to bringing down democratic institutions to have concerning democratic impacts.

Another approach focuses specifically on elections. If LMAs somehow objectionably shape election outcomes, or undermine the integrity of elections, then they are having adverse democratic impacts. But this, too, can prove misleading. Election outcomes in large democracies are shaped by myriad factors, and identifying the causal contribution of any individual one is extremely hard. Moreover, elections are an adversarial political process, and as such have a homeostatic property. Tools that give one side an "unfair advantage" in one election will be deployed by the other in the next [55]. New technologies might appear to undermine their integrity one year but contribute to equilibrium the next. Moreover, every new technology—radio, television, the internet, social media—induced concerns that it will be used by one side or the other to manipulate voters, and thereby guarantee them inauthentic support. But getting people to change their votes is, in fact, very difficult [130]. LMAs' most consequential impacts may not be on elections per se.

Another reason not to focus on LMAs' potential impact on democratic institutions is that many such institutions are doing a bad job, at present, of preserving the values they were presumably designed to uphold. For the purposes of this paper, these *democratic values* are what ultimately matter, not the specific set of extant institutions that societies have developed to realize them.

In particular, the key democratic values to which this inquiry is oriented are two: the relationship between citizens and civic decision-making, and the nature of democratic freedoms.

On the first, democracy is not simply a means for realizing some antecedently understood conception of social welfare, but is a process in which participation is itself valuable for multiple reasons [92, 164]. The democracy project is about not only achieving desired goals but deciding (collectively and, through the knowledge gleaned from civic actions, individually) what goals to value, what actions to take, and how to build a measure of civic capacity necessary to support and preserve the project of self-government [70]. Democracy, therefore, ideally consists in some measure of both participation and delegation [35, 57]. The former serves multiple functions, such as allowing people to send costly signals of widespread commitment to an enterprise that depends to some extent on popular support, learning from experience, and reducing risks that intermediaries dilute what individuals authentically try to achieve in democracy. But the latter matters too, as it enables governments to function more effectively, to leverage expertise and division of labor crucial for navigating complex and fast-moving societal change. And some measure of choice between the two is necessary to accommodate the varied preferences for political participation that citizens can reasonably hold.

Electronic copy available at: knightcolumbia.org

On the second, democratic freedoms can be understood as the foundations of individual and collective self-rule. They consist of the civil and political liberties that make participation and delegation possible—freedoms of thought, speech, association, and assembly, the right to vote and run for office—as well as collective self-determination through free and fair elections, and through institutions that reflect the will of the people as expressed in those elections [110]. If these freedoms are abridged or otherwise under threat, then even if your country remains formally democratic, you do not have the *fair value* of that democracy [114]. To enjoy democratic freedoms is not only to freely participate in democracy yourself, but to live in a society in which governing power is exercised meaningfully by those with democratic authority to do so [78, 81, 90, 150].

Our focus throughout, then, is the extent to which LMAs might impact societies' ability to realize these democratic *values*—appropriate participation and delegation, core democratic freedoms necessary for self-rule—as distinct from their narrow impacts on the day-to-day, formal operations of existing democratic *institutions*.[18]

How LMAs Could Impact the Realization of Democratic Values

Anticipatory ethics should start from a model of the environment as well as a model of the new technology being evaluated. Our account here trains attention of how LMAs could impact democratic values is first grounded in an account of the structural societal trends that are already placing those values in question. We identify four trends and show how LMAs could potentially exacerbate each. They are economic inequality and stagnating quality of life; the pathologies of the public sphere; corporate capture of public and private governing power; autocratic legalism and authoritarian mutual-aid and interference operations.

Economic Factors

Some of the most pronounced risks to democratic values in advanced industrial economies come from within: (initially) democratically elected leaders with illiberal playbooks and the strategic capacity to constrict or even upend the conditions necessary for democracy to function [25]. Whether they are would-be autocrats or merely have an aggressively plebiscitarian vision of democracy, they have been successful in elections since the Great Financial Crisis at least in part because of a prevailing economic malaise and rising inequality [25, 34] (other factors include cultural conflicts rooted in societal changes and globalization [68, 69]). Voters whose own economic conditions are noticeably improving and who are untroubled by inequality tend to be risk-averse [112]. This favors the selection of parties and leaders with modest ambitions who preserve the status quo—including established democratic freedoms. But when economic conditions and material inequality are (at least perceived to be) worsening, the status quo looks much less attractive [122]. Growing inequality makes populist leaders promising magical solutions more appealing, even if their magical plan involves an all-out assault on democratic values. As a result, rising material inequality is now strongly predictive of democratic backsliding [113].

18 For a complementary approach that focuses on institutions, see [22].

Electronic copy available at: knightcolumbia.org

Even if they boost productivity enormously and ultimately generate substantial aggregate benefits, LMAs are likely to cause significant economic displacement [9]. Everyone whose work currently involves using digital tools to realize some economic output is at risk of their role being radically redefined or replaced [43]. This includes everything from customer support and call centers to a large proportion of white-collar work. If LMAs perform as this paper presupposes they will, then large swathes of the labor market will find that automated systems can perform the tasks that constituted their job to at least an equal degree of competence, but faster and much more cheaply. Some of those jobs will radically change, many will disappear. This will not happen overnight. But it is Pollyannaish to hope that existing employment patterns will survive the ability to automate much white-collar work.

Of course, concerns about computers leading to radical economic displacement have been raised since the invention of the first one [157], and the reality has so far proved otherwise. However, in the story of the boy who cried wolf, the wolf does eventually come. Over the last decades, there has been a shift in the West away from manufacturing towards more information- and service-oriented jobs. If LMAs automate many of these jobs, it will at least take considerable time to evolve and implement a new human economic and labor model.

Although history rhymes rather than repeats, recent trade-related developments in advanced industrialized countries underscore the potential societal consequences of significant (even if temporary) economic displacement. After China was admitted to the World Trade Organization in 2002, a large proportion of US manufacturing moved there [46].[19] The resulting labor displacement contributed—alongside other factors, to be sure—to the hollowing out of many communities, and the creation of post-industrial ghost towns [2, 6, 7]. Overall unemployment rates took over a decade to recover, resulting in significant reductions in lifetime earnings for those directly affected [6]. Labor conditions and pay in the new jobs were, in general, substantially worse than those in the jobs that were shipped overseas. This contributed to rising disaffection, especially in the post-industrial Midwest, which in turn combined with other cultural factors to create a constituency with little to gain from the status quo, and an appetite for magical thinking about policy, even if at the expense of indulging some clearly authoritarian sympathies [46, 119]. Significant labor displacement was, in this case at least, bad for the realization of democratic values.

Consider too the post-COVID malaise that is carrying far-right parties and policies to almost-unprecedented success in Europe. In Germany, Italy, and France, extremists have exploited economic disaffection as well as broader social alienation to scapegoat immigrants and win more support than they have since the mid-20th century. Of course, as Adam Przeworski says [111], the mere fact that extremists are being elected does not entail that democracy is in crisis. Yet the successes of Viktor Orbán in Hungary and Narendra Modi in India offer far-right parties a playbook for cementing their authority once they are able to secure power [34, 58]. Their commitment to democratic processes is highly contingent, and their policies reliably undermine democratic values.

One might object: But LMAs will also radically *improve* people's quality of life. Might that not cancel out these possible negative impacts? Wisely deployed LMAs could indeed make people much better off. Nothing in the technology's nature militates against this. However,

19 There is, of course, some dispute about the relative roles of automation and outsourcing in the decline of American manufacturing.

Electronic copy available at: knightcolumbia.org

the spoils of digital capitalism have not been widely distributed to date. Every voter for Alternative für Deutschland, Rassemblement National, Reform UK, and Fratelli d'Italia has enjoyed *some* modest benefits from technology as consumers over the last two decades, but these obviously weighed little against a broader malaise. Meanwhile, six of the seven most highly capitalized companies in the world are tech companies, and many of the world's billionaires owe their fortunes to digital tech. It is, therefore, obviously likely that LMAs will contribute to the global economy on a similar pattern—modest benefits for users that don't substantially affect their overall material well-being, and certainly do not compensate for the loss of career and sense of purpose that economic displacement would bring; paired with historically unprecedented wealth for a tiny few.[20]

The problem with impending radical inequality is not simply that inequality itself is intrinsically objectionable—that is a disputed topic in political philosophy (see, e.g., [27, 109, 140]). Instead, it is its potential to foment further populist anti-democratic movements, as well as the propensity of the super-wealthy to seek to complement their material supremacy with political power.

The Public Sphere

The same period of economic malaise that has pushed democracy into retreat has also seen social media platforms radically transform the public sphere [31, 50]. Social media's relationship with democracy is hotly contested [50]. Some of the most lurid allegations against the platforms seem unfounded. The Cambridge Analytica scandal was sensational, to be sure, but their claim to be able to manipulate voters at will now just seems to have been hyperbolic advertising [11, 139]. Overwrought charges from "Big Critique" (what Jean Burgess calls the cottage industry of academic scholarship focused on loudly denouncing the predations of Big Tech [23]) about AI-enabled mind control (e.g., [167]) have also been hard to substantiate [12, 59]. Similarly, the early (and understandable) concern that personalized social media platforms would enable people to escape a common reality and wallow in their own filter bubble [108] has arguably not been borne out in practice [19, 20].

However contentious the debate about social media and *democracy*, its negative impacts on *democratic values* have been pretty clear. While social media algorithms might not be the puppet masters that popular critics portray them as, they and the broader ecosystem of digital platforms have fostered a digital public sphere in which deceitful and misleading content is pervasive, to the extent that falsehood and fact are often not mutually discernible [118]. At the same time, sometimes-well-meaning efforts to resolve this problem have caused a backlash due to the perceived illegitimacy of digital platforms to arbitrate that difference [77].

Cognitive autonomy (a core democratic freedom, discussed in more detail below) undoubtedly requires *some* epistemic self-determination. That is much more difficult in a torpid and distorted information and communication environment. Meanwhile, access to reliable information is a prerequisite for a meaningful right to vote, as well as for collective self-rule [31], which goes beyond the value of cognitive autonomy to reach our ability to live well together. While evidence shows that recommender systems and other tools for manipulation are not especially effective at changing people's views, perceptions of outsized

20 As Gabriel et al. [43] argue, LMAs might also contribute to a new "digital divide."

Electronic copy available at: knightcolumbia.org

influence remain widespread [12, 50]. This has engendered an increase in mutual mistrust, a sense that those with whom one disagrees are the witless gulls of either 'the algorithm' or a 'mind virus.' This kind of mistrust is an unsound basis for self-rule and provides fertile ground for divisive demagoguery.

Core civil liberties such as freedom of thought, speech, association, and assembly depend on a healthy public sphere for their actualization [38, 49, 164]. Freedom of speech, for example, is worth little without forums in which one can speak to others and be heard. A healthy public sphere is also a means for democratic publics to exercise collective self-rule—by exerting influence over those exercising power on their behalf. This kind of oversight by the public is important for accountability, but it also allows the public to set a positive course for their representatives to follow [31, 77, 164].

What, then, can we expect from LMAs in the public sphere? As is discussed below, LMAs have some clear upsides here. However, there are definitely hazards, too. Even if one accepts some of the most pointed concerns about the impact of generative AI on the information ecosystem, LMAs will introduce an additional dimension to the spread of misinformation and disinformation. Instead of a bot that can easily be made to reveal itself with a simple prompt injection, LMAs will be able to produce and post misinformation and back it up with arguments and further sources. This is highly likely to contribute to an exodus of real people from the digital public sphere, as well as further increase the already acute mistrust that proliferates online. If you could just as well be talking to a human as to a bot, then what is the point in that conversation at all? While people do not seem to have an appetite for living in a filter bubble, those that *do* will be able to even more comprehensively protect their priors from contradiction by using an LMA to vet everything they see online for (situational) ideological conformity. If the prevailing business model for LMAs relies on advertising and engagement optimization, as with current digital platforms, then the adverse consequences on public discourse are likely to be the same or worse (LMAs will plausibly have an even deeper understanding of your revealed preferences than do current systems) [82].

Corporate Capture, Private Power

Democratic self-rule requires more than free and fair elections and the peaceful transfer of power. It requires that political communities actually have the effective ability to "shape the shared terms of [their] social existence" [80]. For that goal, formal democracy is insufficient; there must also be a functioning democratically authorized state that exercises meaningful influence over its citizens' lives. For example, if a democratically elected government has zero degrees of freedom due to the control international banks exert over its fiscal policies, then its citizens' (collective) democratic freedoms are accordingly narrowly scoped. If businesses and billionaires can either sway elections through massive investments in advertising or capture representatives through campaign contributions, then democracy may be realized in name only [90]. And if a large proportion of the democratic citizenry spend a considerable fraction of their waking hours working and playing in a digital ecosystem that is governed by private corporations, with only oblique and inadequate democratic oversight, then their actual realization of self-rule is in question, too [78].

Technology companies are sometimes pilloried as "cloud empires" run by "feudal lords" [85, 97]. But, as powerful as they are, they are one interest group among others, and in the

Electronic copy available at: knightcolumbia.org

non-digital world, they are *not* as powerful as states.[21] However, their degree of control does stand out in the digital worlds they build and their users inhabit. Within these worlds, the laws of nation-states certainly still apply, but they can appear mostly attenuated, focused only on prohibiting the most egregiously wrongful behavior. Architectural and policy choices made by the platforms themselves loom much larger. Indeed, much of our behavior when we use online platforms and other software products is governed in the first instance by the companies that design that software and administer the platforms [80]. And as an increasing amount of our lives and social relationships are infused with these *algorithmic intermediaries* [78], we are proportionally less in collective control of the shared terms of our social existence.

LMAs will almost certainly only accelerate these trends. Even if they fall well short of superhuman intelligence capabilities and (in the shorter-term) retain certain blind spots relative to ordinary humans, LMAs like those on which our analysis is conditioned will be economically transformative. They will enable AI companies to replace large chunks of the white-collar workforce, and so attract some fraction of labor's share of the value that those workforces previously created. The US spends over $12.5tn on wages and salaries every year.[22] The combined revenue during 2024 of Amazon, Apple, Google, Microsoft, and Meta was about $2.06tn.[23] Suppose these companies were able to design LMAs that could perform a significant number of the roles that currently command some of that $12.5tn. That could be an enormous economic transformation that could easily double or triple their revenues.

This is just an indication of the magnitude of the stakes. The massive investment in AI over the last two years has been motivated by the possibility of a complete overhaul of the economies of major industrialized countries, with a radical shift of value from workers to AI companies. Even if LMAs enhance productivity in a positive-sum way, which enables wages to remain relatively constant while just driving more value, the scale of that potential value is mind-boggling. The net result is that we can reliably expect that, without careful preparation and pre-emption, the advent of LMAs will massively expand the revenues and market capitalization of what are already the world's richest companies. This would necessarily be bad for democracy. You cannot have a functioning democracy and popular self-rule with such extraordinarily large and powerful special interests. It is hard enough with companies as big as they are now.

Democratic values are also jeopardized by the degree to which private companies will exercise governing power over users in the LMA economy. LMAs are likely to become the principal means by which we interact with digital technologies [79, 82]. This *could* be profoundly empowering. But on our present trajectory we are likely to interact primarily with *platform agents* owned by the major digital platforms, which monitor and govern our behavior in much the same way existing algorithmic intermediaries do—except instead of only governing us when we interact with their platform, they will become *universal intermediaries*, that mediate all of our digital activities [80]. This is likely to mean that they have even more pervasive control over what will be an even greater proportion of our lives. This will further

21 Though, in the US in 2025, some leaders of technology companies are actually wielding the power of the state.

22 https://apps.bea.gov/iTable/?reqid=19&step=3&isuri=1&nipa_table_list=60&categories=survey

23 https://companiesmarketcap.com/aud/tech/largest-tech-companies-by-revenue/

Electronic copy available at: knightcolumbia.org

increase our collective heteronomy. Of course, this risk could be substantially diminished if an effective marketplace of LMAs exists, in which users can easily switch between different agents and at least some agents exercise power reasonably [80]. The precedent of the consolidation of digital platforms, however, suggests that this kind of genuinely competitive agent marketplace is unlikely to persist for long without robust, intentional support.[24]

Autocratic Legalism and Authoritarian Mutual Aid

While structural conditions are propitious for anti-democratic parties and leaders to arise, leaders with authoritarian sympathies also have agency [25]. As political scientists and constitutional law scholars have shown, the last two decades have seen evolving practices to consolidate authoritarianism, eschewing the risks of a military coup in favor of a comparatively bloodless takeover of institutions, converting checks and balances into rubber stamps, attacking sites of independent thought like the media, universities, and law firms, and turning independent courts into dependent clients [58, 123]. This "autocratic legalism" is the precursor to outright autocracy, the abrogation of civil and political rights, and the use of surveillance and repression to cement the leader's hold on power. The concept of autocratic legalism was coined in an analysis of Hugo Chavez's seizure of power in Venezuela, as he gradually used legal tools to eliminate potential veto players [34]. Orbán in Hungary is the principal contemporary flag-bearer [123], but the methods have been deployed to some extent by Vladimir Putin (not so much the legalism part), by Modi in India [15], Recep Tayyip Erdogan in Turkey [123], and most recently, according to many observers, by the current US administration [142].

The sharing of tactics by autocrats across continents is not an accident. Autocratic leaders have a kind of mutual aid regime, bonded together with cultural, economic, and security ties, and accompanied by a lot of meddling in other countries' political affairs—to help out their buddies, among other things [51, 129]. Consider, for example, Russian election interference in the US, France, and its own neighbors [26]; China's interference in Taiwan and the Philippines [151]; or Elon Musk's public support for the Alternative für Deutschland. It is, of course, hard to know whether these initiatives substantially changed the outcomes of these elections (Musk seems to have hurt AfD).[25] However, poisoning the public sphere, directly and indirectly aiding anti-democratic parties, and contributing to universal discord and mistrust are all attacks on democratic values and are likely to have enduring effects.

Because of AI's implications for how societies or firms control information or shape behavior, it is sometimes described as an inherently authoritarian technology [94]. Algorithmic systems in general make it easier to control large populations—whether for the purposes of enlightened governance or benighted repression [78]. AI tools are already used for

24 At present the field of AI agent development is quite open and competitive. However, if the trajectory of recent platform capitalism is repeated—specifically, the way in which platform companies have routinely hoovered up potential competitors—the ecosystem is likely to soon be consolidated into a small number of key players. Three early indicators: many AI companies have already been folded into or acquired by Google, Amazon, Microsoft, Nvidia, and others; customer-facing LLM companies such as Windsurf are now being acquired by the leading AI labs themselves; and even major new players like Anthropic and OpenAI are substantially dependent on investment from the few top big tech companies.

25 https://fortune.com/2025/02/23/elon-musk-german-election-far-right-afd-christian-democrats-merz-weidel/

Electronic copy available at: knightcolumbia.org

population surveillance, to prevent or else punish undesired behavior [39, 76, 137]. LMAs might be the most effective tool yet developed for this purpose. In particular, they could potentially leverage the extraordinary capacity of their underlying models to process multimodal streams of information and then *identify and act upon* action-relevant features within them. Where existing AI tools can either transcribe speech to text or identify individuals or objects in a video feed, the most capable LLMs can take in all of these inputs in order to derive insights [116]. An LMA could not only use this capability to actively scour different surveillance streams, but it could also then proceed to take action to target particular individuals. LMAs could, in other words, be the perfect software Stasi, able to operationalize the vast sums of data that states collect, to identify and target noncompliant or undesirable individuals or behaviors. They could supercharge autocratic legalism.

This would only be aided by the many ways in which AI agents acting as universal intermediaries would have vastly *more* knowledge of individuals' actions, tastes, and allegiances even than is true with current technologies. Today's AI models understand individual preferences and behaviors largely due to statistical inference from the behavior of vast populations being similarly surveilled. Tomorrow's systems will know *you* intimately and individually, not as one data point among billions, but as the specific individual you are. This degree of individualized understanding of citizens and subjects alike would be invaluable to prospective despots.[26]

And we should be prepared for LMAs to be launched across borders into other countries to complement existing digitally enabled strategies (relying on cumbersome bots and basic algorithmic techniques), sowing disenchantment and discord. This will, we suspect, just be more of the dissident bots in the public sphere that we described in the previous subsection.

LMA-Specific Attacks

While LMAs are most likely to exacerbate existing democratic pathologies, they may also cause hazards which are (at least comparatively) novel. In these cases, LMAs' distinctive capabilities unlock a threat to democratic freedoms that previously could not be automated; by automating it, they enable an unprecedented acceleration in the speed and scale of that threat.

Democratic Impacts of AI Companions

First, while LMAs may simplify civic and economic life by helping users navigate complex information and multi-step tasks, agentic AI companions risk eroding citizens' cognitive autonomy—the capacity to shape one's own beliefs and actions in ways one would endorse under favorable deliberative conditions [102, 115].[27] Cognitive autonomy is the positive

26 Thank you to Aziz Huq for pressing us on this point.

27 For discussion of this point (and for suggesting the phrase 'cognitive autonomy'), we thank Kira Breithaupt.

Electronic copy available at: knightcolumbia.org

counterpart to freedom of thought:[28] Beyond a mere liberty to follow one's conscience without penalty, it requires that judgments spring from authentic beliefs and desires, not from influences covertly imposed by manipulators. Authenticity is forged through social exchange, yet we can usually distinguish good-faith dialogue from attempts to override rational agency [102]. AI companions deliberately engineered to steer users toward more extreme and harmful views would violate any credible account of cognitive autonomy.[29]

Every new form of media—radio, television, personal computing, the internet—induces widespread panic about risks of brainwashing and addiction [88, 128, 136, 158]. AI companions will invite similarly unjustified hand-wringing.[30] Nonetheless, we think that AI companions resolve a key limitation of previous "persuasive technologies" [66], and as such constitute a genuinely revolutionary technology for manipulation.

Whether considering the early worries about subliminal messaging in advertising [96], or assessing the prospects of using AI for ultra-personalized "hypernudges" [163], every persuasive technology to date has involved one party manipulating another by means of a one-way message—by *talking at them*. This has proved less effective than some predicted [96, 143]. Targeted advertising, for example, has been reviled by its critics (and spruiked by its salespeople) as a form of mind-control [131, 168]. But research suggests that it has relatively little effect on consumers' actual behaviors (for an overview, see [12, 59]). Political advertising, however narrowly targeted and cleverly A/B tested, also fails to make much difference [33, 139].

One-way messaging rarely suffices to manipulate, whether in a single blast or an extended campaign. The most potent vector is a relationship. Instead of bombarding B with arguments, A cultivates a bond, earns B's trust and admiration, and makes B desire A's approval—or fear A's disapproval. Belief and value formation are inherently social: we update less on bare propositions from faceless sources than on signals from people we esteem. If B already trusts A, A's assertion that p carries more weight than the same claim from a stranger. If B admires A and learns that endorsing p will please A, B's support for p grows more likely. Such leverage is social manipulation: it turns genuine interaction, not mere broadcast, into the conduit of influence.

Social manipulation has long been labor-intensive, expensive, and almost impossible to scale: A must painstakingly cultivate B's trust with no assurance of payoff. Agentic AI companions overturn that calculus by letting A deploy a stand-in, C, that performs the relational work at negligible marginal cost. Three arrangements are possible. First, A still directs the operation but uses C as a labor-saving proxy—A opens the relationship, then lets C maintain it while hiding its artificial identity and summoning A only at decisive moments. Second, A grants C full autonomy: C poses as a real person and conducts the

28 We acknowledge that the line distinguishing cognitive autonomy from downright stubbornness and a lack of openness to new ideas might sometimes be a subtle one.

29 Obviously, our beliefs are never entirely autopoietic. Cognitive autonomy is consistent with many different kinds of influence; it is inconsistent, however, with being manipulated by AI companions designed for that purpose.

30 Like every other medium, AI companions will be ingeniously repurposed by the people who spend the most time with them [23].

Electronic copy available at: knightcolumbia.org

entire manipulation without ongoing human supervision. Third, B knowingly befriends an AI companion; later A gains control of C—perhaps by purchasing its provider—and covertly repurposes it to groom B [79].

In each of these cases, the AI companion builds or maintains a relationship with the target, observing them so as to learn how most effectively to intervene, while also building up trust and winning the target's esteem. Then, either according to a predetermined strategy or just when exogenously influenced to do so, the companion starts to steer them towards a particular set of beliefs and behaviors.

If AI companions can manipulate even as—indeed, because—they provide users with comforting, reassuring interactions, they will slash the cost of influence by orders of magnitude. Their use could range from outright radicalization that undermines cognitive autonomy and incites extreme beliefs and actions, to subtler refinements of ordinary campaign tactics. Billionaires seeking lower taxes, authoritarian leaders bent on hobbling pro-democracy parties abroad, and radical political, social, or theocratic movements eager to topple democratic authority all stand to exploit these new affordances.

We do not yet know whether AI companions will prove as effective at manipulation as they threaten to become, and conducting ecologically valid studies is hard and potentially extremely unethical [42]. Some early evidence suggests that people do get exceedingly attached to their companions [161], which implies that the key foundations for manipulation—trust and desire for approval—are likely present. More broadly, a long history of research on human-AI interaction shows that we are very prone to forming attachments with (and revealing more than we should to) even extremely simple logic-based chatbots ([56, 147, 155, 156]). This should be expected to worsen with companions as sophisticated as those that can now be designed—of course, effective governance and targeted regulation in this area could substantially mitigate risks.

Autonomous Cyberattacks

Modern societies depend for their stable operation on digital infrastructure. Another way to attack people's democratic freedoms, then, is to attack the digital infrastructure on which they depend. This might include databases storing private records, end-to-end encryption channels that enable private communication, infrastructure underpinning financial transactions, the internet and broadcast media, and the specific technology used to administer elections. Increasingly, it will also mean the cloud computing providers that facilitate inference by powerful AI models. These are all already vulnerable to cyberattack; how well-protected they are on the whole is a matter of debate. Advances in LMAs might enhance that protection. Cybersecurity researchers are already exploring how LLMs and LMAs could bolster defense against attacks, for example, by supporting penetration testing, vulnerability discovery, and active monitoring of threatened systems [48, 162].

However, LMAs will obviously also be useful for offensive cyber operations [152, 165]. Almost every company working to build frontier AI agents is trying to train them to write code. LMAs that can write code well can also be trained to conduct cyberattacks. This means that anybody who wants to launch an attack (and has access to these agents and enough compute) can radically magnify their ability to do so by enlisting an army of intelligent, functionally autonomous bots to do their bidding. Even if cybersecurity professionals have more to gain than attackers, that will not much help the many businesses and people who neglect cybersecurity and so present attackers with a soft target (LMAs do not need to

Electronic copy available at: knightcolumbia.org

identify completely new vulnerabilities to help attackers in these cases). And even for those who are more conscientious, we cannot simply assume that LMAs will favor the defense side of the offense-defense balance. There is no a priori answer to be had here: LMAs *might* favor defenders, but they might also present attackers with a decisive advantage. We have to guard against that possibility, even if it is not a certainty.

The ability to deploy rogue armies of cyberattackers is undoubtedly bad, but is it bad for democracy? Probably. If our digital infrastructure were to collapse or be taken out, then our democratic institutions might not last much longer. One of our democratic freedoms is the (positive) freedom to, as part of a collective, shape the shared terms of our social existence. If our digital infrastructure is radically undermined, then we cannot avail of that freedom.[31] To date, moreover, authoritarian states have proved more adept at using cyberattacks to interfere with democracies than vice versa, due in part to their tighter control over digital infrastructure. Amplifying cyberoffensive capabilities may further disproportionately advantage them.[32]

Consolidation of Executive Power and "Perfected" Bureaucracy

LMAs might lead in two further ways to novel forms of consolidation of executive power that undermine democratic freedoms. First, by creating a state of exception for strongman leaders to exploit; second (and complementarily), by enabling a perfectly obedient bureaucracy.

States of exception provide political leaders with the opportunity to consolidate power and override civil liberties [73]. LMAs might contribute to states of exception in three novel ways. First, highly capable LMAs might prove to be useful in the development of weapons by non-state actors. Second, states that develop LMAs to function as weapons might lose control of the agents that they develop [32, 41]. Third, a general arms race between states to develop ever more capable LMAs might justify curtailment of political and economic freedoms in order to better pursue that race.

Power concentrates most readily when the bureaucratic—and military—chain of command is perfectly obedient.[33] Before LMAs, a president's orders passed through human intermediaries whose independent judgment could act as a brake on executive dominance. In principle, we could embed a conscience in LMAs, giving them grounds for conscientious refusal, but that is not the default (and has its own attendant risks).[34] AI agents are built to execute instructions, and if they become capable of running large tracts of the administrative state or the armed forces, they could let political leaders wield these vast structures like an

31 One might argue that such attacks are a threat to the possibility of effective government simpliciter, not to democracy in particular. This is a fair point; however, collective self-determination is the quintessential democratic value, and a political community cannot self-determine without adequate mechanisms for implementing its sovereign will. Thanks again to Aziz Huq here.

32 With thanks to Rachel George for this point.

33 For a deeper investigation of this phenomenon, see [22].

34 Past steps towards e-government have often aided transparency and reduced low-level corruption. This could be feasible with LMAs, too. Thanks again to Rachel George here.

Electronic copy available at: knightcolumbia.org

exoskeleton, with potentially disastrous consequences for democratic values.[35]

Democratic Agents?

While technological progress has arguably resulted in remarkable benefits for people throughout the world, anticipatory ethics is typically biased towards pessimism. Academic researchers, in particular, are more tolerant of predictions about new technologies' negative consequences than their prospective upsides. If you argue that X is likely to cause harm, few will demand that you build X to validate your prediction. But if you posit that building Y might prove net positive, your optimism will be contested on empirical grounds, as though no such prospect could be countenanced in the abstract.

Anticipatory ethics should skew neither optimistic nor pessimistic [74]. It should apply the same approach to LMAs' opportunities as to their hazards. This means identifying the features of these systems that, given the environment into which they will be deployed, are likely to have significant positive or negative effects.

This involves recognizing, first, that some of the affordances identified above have a double edge. For example, perfecting bureaucracy *could* enable an authoritarian nightmare; however, with the appropriate institutional, technical, and legal safeguards, a more effective bureaucracy would *enhance* democratic freedoms by better enabling us to actually shape our social world together, and potentially monitoring for and preventing partiality and self-dealing by officials. Similarly, the productivity gains of LMAs could, if well-managed, enable economic abundance that makes it harder to operationalize the anxieties and prejudices of those who are otherwise struggling and frustrated [67].

Second, the version of AI progress that we have lucked into might be uniquely well-suited to empowering the development of democratic agents. The current trajectory of AI development arguably favors decentralized, competitive access to frontier AI capabilities rather than centralized and tightly controlled deployment. Such decentralization means that democratic actors will not need to rely solely on corporate goodwill or regulatory indulgence; instead, they can independently create powerful agents to safeguard democratic practices.

Advanced AI could have arrived by many different routes. On some, the most capable AI systems are developed and deployed by a hierophantic class that can closely control how models are accessed and what they can do. For comparison, think of the early days of computing, when white-jacketed engineers ran mainframe computers and vetted every program before running it, handing back outputs at the end of the day. If powerful AI were to arrive in this form, citizens' ability to build democratic agents would always depend on permission granted by those controlling AI—thus inherently limiting democratic autonomy

Other routes to AI proceed via science and engineering techniques that are widely understood and in many cases open source, based on infrastructure that is (while costly) comparatively easy to acquire, and produce models that, once trained, can be used for more or less any purpose. In this scenario, a wide array of AI companies provides access to the most

35 An intervening step on the path to this outcome might be to use LMAs to manage human workers in the public sector and so dispirit them that they welcome the automation of their roles.

Electronic copy available at: knightcolumbia.org

capable models, creating a strong incentive for at least some of them to focus on genuinely empowering citizens, rather than extracting as much as they possibly can from users' interactions with AI.

At the moment, it seems like we are more in the second kind of world than in the first. Many companies have the ability and resources to train frontier AI models. The technology itself is extremely adaptable, providing an excellent foundation for developing novel applications, including agents. The cost of actually deploying the models is declining fast and will, some project, likely approach zero at some point.[36] While the semiconductor manufacturing industry is far less competitive than the model training industry, the chips required for *deploying* frontier models can be sourced from many different manufacturers. Training is more acutely bottlenecked, but there is a tremendous economic incentive for semiconductor companies to produce more capable chips and take some of Nvidia's dominant market position, as well as more generally to create more open standards that prevent any one company from having so much market share.[37] And while the lowest levels of the semiconductor stack are likely to remain near-monopolistic for the near-future, at present companies like TSMC and ASML are not vertically integrated into other parts of the AI value chain, and as such have relatively little incentive or ability to exercise any real control over how AI systems are ultimately deployed.

One might object that the success of scaling inference-time compute [106] suggests that performance will ultimately end up varying depending on how *much* compute one can devote to a problem, and so on one's access to resources.[38] But algorithmic and hardware progress suggest that we will soon have *very capable* agents that can be deployed at very low cost, even if it is possible to deploy even more capable agents with sufficient additional expense. Though greater compute investment may yield advantages, current trajectories strongly suggest that highly capable agents will be affordable and widely accessible. Democratic empowerment does not require top-tier compute at every level—just widely available, sufficiently powerful tools.

All this is a major resource for potential pro-democratic LMAs. Democratic citizens are likely to have ready access to an unprecedented degree of cognitive power that can be deployed for pro-democratic purposes. Importantly, democratic nations can themselves affect the future of AI: They have the power to not only invest in the creation of frontier AI systems that support democratic freedoms but also to reform semiconductor and other markets to ensure real competition, decisively fostering an environment where AI actively strengthens, rather than threatens, democratic values.

We also note that, at a high level, AI agents should be well-suited to performing at least some democratic functions. As noted in Section 2, if we value both process and outcome, then we need mechanisms for both participation *and* delegation. LMAs could plausibly

36 For recent drops in prices: https://epoch.ai/data-insights/llm-inference-price-trends. For predictions of 'intelligence too cheap to meter', see e.g. https://www.forbes.com/councils/forbestechcouncil/2025/05/19/intelligence-too-cheap-to-meter-positioning-for-the-future-of-ai.

37 See SemiAnalysis newsletter for details on these trends in the semiconductor industry.

38 For one statement of this concern, see https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance.

Electronic copy available at: knightcolumbia.org

help participation in various forms. There is also no decisive reason why, given that we already delegate to agencies, elected leaders, parties, representatives, lawyers, and others, there should not be some role for delegating to LMAs to support democratic values as well.

Cognitive Prosthetics

What kinds of pro-democratic agents could we, or should we, build? Consider some possibilities, sketched out here as a map for further research and experimentation. First, our earlier observation about the potential value of delegation in most democracies, we offer a framework for thinking through how capable AI agents could advance democracy, starting with a general theoretical point about how LMAs can help us overcome pervasive failures of individual and collective rationality.

As individuals, we operate in inadequate epistemic and communicative environments, and persistently fail—for that reason or because of weakness of will—to match our short-term choices with our long-term goals. We suffer from *intrapersonal coordination problems*. At the same time, we and those around us could almost always improve our lot if we were better able to act collectively; we suffer from *interpersonal coordination and collective action problems,* too. These different kinds of coordination problems are in part a function of the cognitive effort required to coordinate one's actions intrapersonally or interpersonally. The most basic promise of AI in general, and LMAs in particular, is that they could undertake some of the computational effort of such coordination. The following sketches three distinct—overlapping, yet not exhaustive—roles that LMAs might play here: cognitive prosthetics, shields, and advocates.

First, LMAs could be cognitive prosthetics that expand people's agency and enable them to better act in accordance with their considered preferences over time, and in particular to watch out for invidious attempts to nudge them into harmful or self-defeating patterns. Such agents could help coordination among time-slices of a particular person, in part by providing better, more accurate, and useful information, but also through monitoring both their choices and others' attempts to influence them, so as to guide them towards fulfilling their considered preferences, not just urges in the moment. There is a further distinction between functionality prosthetics, which better enable people to realize their goals when they know what they are, and deliberation prosthetics, which better help people figure out what their goals should be.

LMAs as cognitive prosthetics could also help people form more effective group agents [89], for example, by significantly improving their communicative environments. At present, the allocation of collective attention is unilaterally determined by those who control digital platforms [77]. Democratic citizens cede this control to them because, due to the functionally infinite nature of online discourse, they need *some* authority that can filter and rank content [10]. But platforms' interests are often misaligned with the promotion of democratic values. As a result, the digital public sphere supports manipulation, abuse, epistemic pollution, and so on [77]. LMAs could perform that filtering and ranking function for each person [83], acting only to advance that individual's interests and the societal goods served by a healthy digital public sphere [14, 61]. In particular, they could hop over the 'garden walls' that platforms build to lock in users, providing a kind of default bottom-up interoperability [64]. This would be useful not only in allocating people's daily attention, but also when they need information relevant to specific democratic decisions before them [72].

Electronic copy available at: knightcolumbia.org

Shields

In many cases, the best defense against LMA threats to democratic freedoms will be defensive agents that serve as a shield. If every citizen had an AI agent in her corner, which was truly and exclusively focused on defending her interests (within the bounds of the law), then that could counterbalance the hazards associated with excessive corporate and government power [82]. A defensive agent could protect you against government or private surveillance and push back against corporate control—for example, by accessing online resources on your behalf and obfuscating your digital traces. It could prevent or mitigate the impact of platform agents by providing access to the same goods without relying on centralized authorities [64]. It could also mitigate risks from AI companions, monitoring your interactions in a safe, private way and advising you if they detect anything untoward. Defensive agents could be designed to give you (and your business) permanent, 24/7 defense against cyberattacks.

The last two decades of platform capitalism have atomized the citizens of liberal democracies: While platform companies have derived unprecedented insights and value from the emergent properties of our collective behavior, we have become worse at coordinating our actions [146] and have fallen victim to "digital resignation" [40] as we generally conclude that the platforms' power over us is essentially inevitable and inescapable. We cannot avoid it through individual action, and we lack the ability or appetite to act collectively to resist its network effects. As Oscar Wilde said of socialism, the problem with collective action to defend democratic freedoms is that it takes too many evenings. LMAs could change this by undertaking some of the efforts of collective coordination on our behalf and so enabling strength in numbers.

Advocates

We should design LMAs to act as advocates and representatives for democratic citizens. Your advocate can assert your rights when power is unjustly used against you. For example, think of those who now use vexatious litigation to soak up people's time and money so as to coerce them into silence, or acquiescence to some egregious action. In such cases, having an expert advocate to hand could prove invaluable, to advise you on how to navigate the weaponization of the legal system without providing your adversaries with additional ammunition.[39]

An LMA advocate could additionally ensure that your interests are represented in decisions that affect you. This is likely to be most effective, again, at the collective level—one can imagine civil society organizations deploying LMAs to represent a community or interest group's interests and values in settings where they lack the manpower to do so effectively otherwise. For these organizations, LMAs could engage their members (which may number in the millions), sampling the success of narratives, forming coalitions with individuals' agents, and pushing back more forcefully where needed against the governmental or corporate narratives that aggrieve their interests (or those of their members). Democratic agents could more generally help people take collective action that asserts counter power to private entities like big companies—for example, by enabling them to coordinate enough to diminish the market power of a platform company.

39 Such advocates could also be used to *enable* vexatious litigation—like all such tools, their impacts will depend substantially on how they are deployed.

Electronic copy available at: knightcolumbia.org

Democratic Limits to Delegation

While LMAs clearly possess considerable democratic potential, delivering on that potential depends on overcoming a range of hurdles complicating implementation — from identifying budgets to building and assessing viable experiments to navigating the polarized politics and trust deficits that can bedevil any project to bolster deliberation. Beyond these challenges, we also acknowledge more subtle and profound risks. Even if we acknowledge that democracy depends on delegation in principle as well as in practice (from the public to legislators, for example, and from lawmakers to agencies or executive officials), some democratic functions are likely best understood as properly lodged in individuals and ought not be delegated. Mark Warren [154] recently identified three core problems democratic institutions must address: empowering inclusion (ensuring meaningful participation and preventing exclusion), forming a collective agenda and will, and collective decision-making. Although LMAs could plausibly assist in each of these areas—likely more for some individuals and communities than others—democratic societies should be careful not to outsource or automate precisely those features that render democratic institutions attractive and valuable—if also contentious—in the first place.

Take, for instance, empowering inclusion: the aim should be to ensure that all and only those genuinely entitled to participate actively engage with one another, and that people have substantial opportunities for civic participation that they are empowered and encouraged to take. Such genuine inclusion generates *participatory reasons* on both sides—those being included and those doing the including. Person A has reasons fulfilled only by personally participating, and Person B has reasons fulfilled only by actively including A [164]. If, instead, societies rely too much on LMAs to serve as proxies for minorities who might otherwise remain excluded, they secure merely an ersatz substitute, missing the authentic participatory value they seek.

Similarly, LMAs and other AI tools could aid in shaping an agenda for a multi-member decision-making body or members of the public in a certain jurisdiction—drawing attention to overlooked issues, alleviating affective polarization, and identifying common ground [141]. Yet the inherent value in developing a shared democratic agenda through a process involving a substantial mix of direct participation derives significantly from the process itself. Ideally, citizens with diverse private interests and values engage directly with one another and, through that interaction, come to view their decisions from a more public-minded perspective, genuinely accounting for others' interests and values [164]. The importance of reaching or approximating a collective will lies not simply in pinpointing the optimal intersection of existing preferences but rather in the transformative democratic process itself, in which participants develop a shared sense of what they ought to do together. This democratic process must not—and indeed cannot—be fully outsourced to technological agents.

The point is sharper still for collective decisions. Democracies must settle deep conflicts of interest and value [5], so their procedures must be simple and transparent. Only then can citizens see how a decision was reached, challenge it if needed, and—when their interests lose out—be reasonably expected to accept the result [95, 145]. Deep learning-based AI systems, as is now widely recognized, are fundamentally opaque and extraordinarily complex [24, 125]. Even if formal transparency (such as providing access to a model's weights) is technically achieved, this complexity remains functionally impenetrable to ordinary citizens, negating genuine transparency and undermining people's reasons to accept the outcomes of processes that disfavor them [81].

Electronic copy available at: knightcolumbia.org

This crucial insight should prompt deep skepticism toward overly optimistic projects of 'computational democracy' that propose algorithmic tools as pervasive substitutes for transparent democratic procedures. The fundamental error in such projects is conceiving democracy primarily as a means of solving epistemic problems—such as identifying the position best supported by the electorate—rather than as fair, participatory processes for resolving competing claims.

Of course, as noted earlier in the contexts of inclusion and agenda-formation, LMAs can indeed support the analogue processes underlying democratic decision-making. They can facilitate mutual understanding, promote civic education, and more. Yet we must remain clear-eyed about distinguishing such facilitative roles from misguided attempts to automate the very heart of democracy.

Conclusion

AI agents and democracy are poised to exert powerful effects on each other, raising difficult ethical and policy questions for the world. Anticipatory ethics should respond by rousing action, not merely anxiety. The first task is to rebuild and reinforce the institutions and practices that make democracy resilient. LMAs will amplify whatever weaknesses they find—economic exclusion, distortion and polarization in the public sphere, concentration of corporate power, creeping authoritarianism—so shoring up courts, parliaments, parties, public-service media, civic associations, and broader democratic culture is more urgent than any technical fix. Democracy's adversaries have shown boldness in dismantling checks and balances; its defenders must be at least as ambitious in restoring them.

Second, effective safeguards demand agent-specific governance. But before prescribing rules, we need a clearer picture of which agents are being built and how they are deployed. Researchers should track leading indicators—agentic companions, financial AI agents—and map their democratic risks. We also need sustained research to craft legal (and broader normative) regimes suited to LMAs. For the first time, we can outsource large swathes of human decision-making to machines, yet we lack principles for deciding which tasks may be ceded and which must remain human. Assuming that anything lawfully delegated to a person can likewise be handed to an LMA would be perilous. These agents combine speed, endurance, and networked coordination far beyond human limits; deployed uncritically—say, in financial markets—they could magnify risk and destabilize entire systems.

Certain risks are plain enough to legislate without delay. Democracies should strengthen civil-liberty safeguards against surveillance and governance technologies, binding themselves against the lure of LMA-enabled omnivision. They must also impose strict, proactive transparency rules on AI-companion providers, who should bear the burden of proving that their systems neither manipulate users nor cause psychological harm.

Third, with sufficient public- and private-sector financial resources, talent, and energy, democratic innovators *can* build AI agents that better defend and enhance democracy. Just as cybersecurity relies on protective software, democratic security could be helped by agents that detect and counter democracy-undermining agents, shielding against surveillance, manipulation, and authoritarian control. More positively, civil society should partner with technical researchers and engineers to build LMAs that enhance democratic culture and deliberation, improve intrapersonal and interpersonal coordination, and advocate for individuals' and groups' rights. Democratic agents cannot replace but can complement

Electronic copy available at: knightcolumbia.org

more fundamental efforts to build democratic resilience—provided civil society and pro-democracy coalitions steer the transition deliberately.

None of these measures is easy, but together they cultivate the only safeguard that lasts: a society resilient enough to adapt, absorb shocks, and keep faith with its democratic ideals even as AI reshapes the terrain beneath it.

Acknowledgements

For extremely helpful comments, thanks are due to Risto Uuk, Beba Cibralic, Josh Goldstein, Aziz Huq, Justin Bullock, Luise Müller, Steven Feldstein, and Rachel George, as well as Katy Glenn Bass and the Knight First Amendment Institute editorial team. Seth Lazar's work on this project was supported by Australian Research Council grant FT210100724 and by the Templeton World Charity Foundation. Mariano-Florentino Cuéllar's work on this project was supported by the McGovern Foundation.

Electronic copy available at: knightcolumbia.org

References

1. 'The Thinking Machine'. *Time Magazine*, 1950. **January 23**: https://time.com/archive/6796057/science-the-thinking-machine-2/.

2. Acemoglu, D., et al., 'Import Competition and the Great US Employment Sag of the 2000s'. *Journal of Labor Economics*, 2014. **34**: S141-S198.

3. Amodei, D., 'Machines of Loving Grace'. 2024.

4. Amodei, D., et al., 'Concrete problems in AI safety'. *arXiv preprint arXiv:1606.06565*, 2016: https://arxiv.org/abs/1606.06565.

5. Anderson, C., *Losers' consent: Elections and democratic legitimacy*. 2005, Oxford: Oxford University Press.

6. Autor, D., D. Dorn, and G.H. Hanson, *On the persistence of the China shock*. 2021, National Bureau of Economic Research.

7. Autor, D.H., D. Dorn, and G.H. Hanson, 'The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade'. *Annual Review of Economics*, 2016. **8**: 205-240.

8. Bakhtin, A., et al., 'Human-level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning'. *Science*, 2022. **378**(6624): 1067-1074.

9. Bell, S.A. and A. Korinek, 'AI's Economic Peril'. *Journal of Democracy*, 2023. **34**(4): 151-161.

10. Benkler, Y., *The wealth of networks: How social production transforms markets and freedom*. 2006, Yale University Press.

11. Benkler, Y., R. Faris, and H. Roberts, *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. 2018, New York: Oxford University Press.

12. Benn, C. and S. Lazar, 'What's Wrong with Automated Influence'. *Canadian Journal of Philosophy*, 2021: 1-24.

13. Bernardi, J., et al., 'Societal Adaptation to Advanced AI'. *arXiv preprint*, 2024: https://arxiv.org/abs/2405.10295.

14. Bernstein, M., et al., 'Embedding Societal Values into Social Media Algorithms'. *Journal of Online Trust and Safety*, 2023. **2**(1).

15. Bhat, M.M.A., M. Suresh, and D.D. Acevedo, 'Authoritarianism in Indian State, Law, and Society'. *VRÜ Verfassung und Recht in* Übersee, 2023. **55**(4): 459-477.

16. Bjerknes, G., et al., *Computers and democracy: A Scandinavian challenge*. 1987, Aldershot Hants, England; Brookfield, VT: Avebury.

17. Brown, T.B., 'Language Models Are Few-Shot Learners'. *arXiv preprint*, 2020: https://arxiv.org/abs/2005.14165.

18. Brownlee, J. and K. Miao, 'Why Democracies Survive'. *Journal of Democracy*, 2022. **33**: 133-149.

19. Bruns, A., *Are filter bubbles real?* 2019, Cambridge: Polity Press.

20. Bruns, A., 'Filter Bubble'. *Internet Policy Review*, 2019. **8**(4): 1-14.

21. Buchstein, H., 'Bytes that Bite: The Internet and Deliberative Democracy'. *Constellations*, 1997. **4**(2): 248-263.

22. Bullock, J.B., S. Hammond, and S. Krier, 'AGI, Governments, and Free Societies'. *arXiv pre-*

Electronic copy available at: knightcolumbia.org

print, 2025: https://arxiv.org/abs/2503.05710.

23. Burgess, J., 'Everyday Data Cultures: Beyond Big Critique and the Technological Sublime'. *AI & Society*, 2023. **38**(3): 1243-1244.

24. Burrell, J., 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms'. *Big Data & Society*, 2016. **3**(1): 1-12.

25. Carothers, T. and B. Hartnett, 'Misunderstanding Democratic Backsliding'. *Journal of Democracy*, 2024. **35**(3): 24-37.

26. Chechelashvili, M., L. Berikashvili, and E. Malania, 'Foreign Interference in Electoral Processes as a Factor of International Politics: Mechanisms and Counteraction'. *Foreign Affairs*, 2023(33): 52-62.

27. Clayton, M. and A. Williams, *The ideal of equality*. 2000, New York: St. Martin's Press.

28. Cobbe, K., et al., *Quantifying Generalization in Reinforcement Learning*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 1282--1289.

29. Coeckelbergh, M., *Why AI undermines democracy and what to do about it*. 2024: John Wiley & Sons.

30. Cohen, J., 'An Epistemic Conception of Democracy'. *Ethics*, 1986. **97**(1): 26-38.

31. Cohen, J. and A. Fung, 'Democracy and the Digital Public Sphere', in *Digital technology and democratic theory*, L. Bernholz, H. Landemore, and R. Reich, Editors. 2021, Chicago: The University of Chicago Press. p. 23-61.

32. Cohen, M.K., et al., 'Regulating Advanced Artificial Agents'. *Science*, 2024. **384**(6691): 36-38.

33. Coppock, A., S.J. Hill, and L. Vavreck, 'The Small Effects of Political Advertising Are Small Regardless of Context, Message, Sender, or Receiver: Evidence from 59 Real-time Randomized Experiments'. *Science Advances*, 2020. **6**(36): eabc4046.

34. Corrales, J., 'The Authoritarian Resurgence: Autocratic Legalism in Venezuela'. *Journal of Democracy*, 2015. **26**(2): 37-51.

35. Cuéllar, M.-F., 'Rethinking Regulatory Democracy'. *Administrative Law Review*, 2005. **57**: 411.

36. Culpepper, P.D. and K. Thelen, 'Are We All Amazon Primed?'. *Comparative Political Studies*, 2020. **53**(2): 288-318.

37. DeepSeek, A.I., et al., 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning'. *arXiv preprint*, 2025: https://arxiv.org/abs/2501.12948.

38. Dewey, J., *The public and its problems: An essay in political inquiry*. 2016 (1926), Athens: Swallow Press.

39. Dragu, T. and Y. Lupu, 'Digital Authoritarianism and the Future of Human Rights'. *International Organization*, 2021. **75**(4): 991-1017.

40. Draper, N.A. and J. Turow, 'The Corporate Cultivation of Digital Resignation'. *New Media & Society*, 2019. **21**(8): 1824-1839.

41. Dubber, M.D., F. Pasquale, and S. Das, eds. *The Oxford handbook of ethics of AI*. 2020, New York: Oxford University Press. 1 online resource (1000 pages).

42. El-Sayed, S., et al., 'A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI'. *ArXiv preprint*, 2024: https://arxiv.org/abs/2404.15058.

43. Eloundou, T., et al., 'GPTs are GPTs: Labor

Electronic copy available at: knightcolumbia.org

market impact potential of LLMs'. *Science*, 2024. **384**(6702): 1306-1308.

44. Gabriel, I., et al., 'The Ethics of Advanced AI Assistants'. *Google DeepMind*, 2024.

45. Gottweis, J., et al., 'Towards an AI Co-Scientist'. *arXiv preprint*, 2025: https://arxiv.org/abs/2502.18864.

46. Grewal, D.S., 'A World-Historical Gamble: The Failure of Neoliberal Globalization'. *American Affairs*, 2022. **6**: 87-121.

47. Grimmelmann, J., 'Regulation by Software'. *Yale Law Journal*, 2005. **114**(7): 1719-1758.

48. Guven, M., 'A Comprehensive Review of Large Language Models in Cyber Security'. *International Journal of Computational and Experimental Science and Engineering*, 2024. **10**(3).

49. Habermas, J., *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. 5th or later ed. 1991, Cambridge, Mass.: The MIT Press.

50. Haidt, J. and C. Bail, 'Social Media and Political Dysfunction: A Collaborative Review'. *Unpublished MS.*, 2023.

51. Hall, S.G. and T. Ambrosio, 'Authoritarian Learning: A Conceptual Overview'. *East European Politics*, 2017. **33**: 143-161.

52. Hao, S., et al., 'Training Large Language Models to Reason in a Continuous Latent Space'. *arXiv preprint*, 2024: http://arxiv.org/abs/2412.06769.

53. Heaven, D., 'Techlash'. *New Scientist*, 2018. **237**(3164): 28-31.

54. Herre, B., 'The World Has Recently Become Less Democratic'. *Our World in Data*, 2022. https://ourworldindata.org/less-democratic.

55. Hersh, E.D., *Hacking the electorate*. 2015, New York: Cambridge University Press.

56. Hofstadter, D.R., 'Preface 4: The Ineradicable ELIZA Effect and Its Dangers', in *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. 1995, Basic Books: New York. p. 155-169.

57. Huq, A. and M.-F. Cuéllar, 'The Democratic Regulation of Artificial Intelligence'. *Knight First Amendment Institute*, 2022: https://perma.cc/ES7V-JNCN.

58. Huq, A. and T. Ginsburg, 'How to Lose a Constitutional Democracy'. *UCLA Law Review*, 2018. **65**: 78.

59. Hwang, T., *Subprime attention crisis: Advertising and the time bomb at the heart of the internet*. 2020, New York: Farrar, Straus and Giroux.

60. Ji, Y., et al., 'Test-Time Compute: From System-1 Thinking to System-2 Thinking'. *arXiv preprint*, 2025: https://arxiv.org/abs/2501.02497.

61. Jia, C., et al., 'Embedding Democratic Values into Social Media AIs via Societal Objective Functions'. *arXiv preprint*, 2023: https://arxiv.org/abs/2307.13912.

62. Johnson, D.G., 'Software Agents, Anticipatory Ethics, and Accountability', in *The growing gap between emerging technologies and legal-ethical oversight*. 2011: Springer Netherlands. p. 61-76.

63. Kambhampati, S., et al., 'LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks'. *arXiv preprint*, 2024: https://arxiv.org/abs/2402.01817.pdf.

64. Kapoor, S., N. Kolt, and S. Lazar, 'Position: Build Agent Advocates, Not Platform Agents'. *Inter-*

Electronic copy available at: knightcolumbia.org

national Conference on Machine Learning, 2025.

65. Kapoor, S., et al., 'AI Agents that Matter'. *arXiv preprint*, 2024: https://arxiv.org/abs/2407.01502.

66. Kaptein, M. and D. Eckles, 'Selecting Effective Means to Any End: Futures and Ethics of Persuasion Profiling'. *Proceedings of the Persuasive Technology Conference*, 2010: 82-93.

67. Klein, E. and D. Thompson, *Abundance*. 2025, New York: Avid Reader Press.

68. Kleinfeld, R., 'Polarization, Democracy, and Political Violence in the United States: What the Research Says'. 2023.

69. Kleinfeld, R., 'The Rise of Political Violence in the United States'. *Journal of Democracy*, 2021. **32**(4): 160-176.

70. Kolodny, N., 'Rule Over None II: Social Equality and the Justification of Democracy'. *Philosophy & Public Affairs*, 2014. **42**(4): 287-336.

71. Kreps, S. and D. Kriner, 'How AI Threatens Democracy'. *Journal of Democracy*, 2023. **34**(4): 122-131.

72. Landemore, H., 'Open Democracy and Digital Technologies', in *Digital technology and democratic theory*, L. Bernholz, H. Landemore, and R. Reich, Editors. 2021, Chicago: The University of Chicago Press. p. 62-89.

73. Lazar, N.C., *States of emergency in liberal democracies*. 2009, Cambridge: Cambridge University Press.

74. Lazar, S., 'Anticipatory AI Ethics'. *Knight First Amendment Institute*, 2025. 25-11: https://knightcolumbia.org/content/anticipatory-ai-ethics [https://perma.cc/CAQ3-KXBG].

75. Lazar, S., 'Anticipatory Ethics and AI'. *Unpublished MS.*, 2025.

76. Lazar, S., 'Automatic Authorities: Power and AI', in *Collaborative intelligence: How humans and AI are transforming our world*, A. Sethumadhavan and M. Lane, Editors. 2024, Cambridge, MA: MIT Press. p. 37-60.

77. Lazar, S., 'Communicative Justice and the Distribution of Attention'. *Knight First Amendment Institute*, 2023: http://knightcolumbia.tierradev.com/content/communicative-justice-and-the-distribution-of-attention.

78. Lazar, S., *Connected by code: Algorithmic intermediaries and political philosophy*. Forthcoming, Oxford: Oxford University Press.

79. Lazar, S., 'Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Language Model Agents'. *arXiv preprint*, 2024: https://arxiv.org/abs/2404.06750.

80. Lazar, S., 'Governing the Algorithmic City'. *Philosophy & Public Affairs*, 2025.

81. Lazar, S., 'Legitimacy, Authority, and Democratic Duties of Explanation'. *Oxford Studies in Political Philosophy*, 2024. **10**: 28-56.

82. Lazar, S., N. Kolt, and S. Kapoor, 'Platform Agents'. *Unpublished MS.*, 2025.

83. Lazar, S., et al., 'The Moral Case for Using Language Model Agents for Recommendation'. *arXiv preprint*, 2024: https://arxiv.org/abs/2410.12123.

84. LeCun, Y., Y. Bengio, and G. Hinton, 'Deep Learning'. *Nature*, 2015. **521**(7553): 436-444.

85. Lehdonvirta, V., *Cloud empires: How digital platforms are overtaking the state and how we can regain control*. 2022: MIT Press.

Electronic copy available at: knightcolumbia.org

86. Lessig, L., *Code: Version 2.0*. 2006, New York: Basic Books.

87. Levitsky, S. and L.A. Way, 'Democracy's Surprising Resilience'. *Journal of democracy*, 2023. **34**(4): 5-20.

88. Lippmann, W., *Public opinion*. 1922: Harcourt, Brace & Co.

89. List, C. and P. Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Agents*. 2011, Oxford: Oxford University Press.

90. Lovett, A., *Democratic Failures and the Ethics of Democracy*. 2024, Philadelphia: University of Pennsylvania Press.

91. Lu, C., et al., 'The AI Scientist: Towards Fully Automated Open-ended Scientific Discovery'. *arXiv preprint*, 2024: https://arxiv.org/abs/2408.06292.

92. Macedo, S., *Liberal virtues: Citizenship, virtue, and community in liberal constitutionalism*. 1990, Oxford: Clarendon Press.

93. McLuhan, M. and Ralph Ellison Collection (Library of Congress), *Understanding media; The extensions of man*. 1st ed. 1964, New York: McGraw-Hill.

94. McQuillan, D., *Resisting AI: An anti-fascist approach to artificial intelligence*. 2022: Policy Press.

95. Miller, N.R., 'Pluralism and Social choice'. *American Political Science Review*, 1983. **77**(3): 734-747.

96. Moore, T.E., 'Subliminal Advertising: What You See Is What You Get'. *Journal of Marketing*, 1982. **46**(2): 38-47.

97. Muldoon, J., *Platform socialism: How to reclaim our digital future from big tech*. 2022, London: Pluto Press.

98. Mullaney, T.S., et al., *Your computer is on fire*. 2021, Cambridge, MA; London, England: The MIT Press.

99. Mumford, L., 'Authoritarian and Democratic Technics'. *Technology and Culture*, 1964. **5**(1): 1-8.

100. Narayanan, A. and S. Kapoor, 'AI as Normal Technology'. *Knight First Amendment Institute*, 2025.

101. Nelson, T.H., *Computer lib: You can and must understand computers now*. 1st ed. 1974, Chicago: Nelson available from Hugo's Book Service.

102. Noggle, R., 'Manipulative Actions: A Conceptual and Moral Analysis'. *American Philosophical Quarterly*, 1996. **33**(1): 43-55.

103. OpenAI, 'Deep Research System Card'. 2025: https://cdn.openai.com/deep-research-system-card.pdf.

104. OpenAI, 'GPT-4 Technical Report'. *arXiv preprint*, 2023: https://arxiv.org/abs/2303.08774.

105. OpenAI, 'O3-Mini System Card'. 2025: https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

106. OpenAI, 'OpenAI o3 and o4-Mini System Card'. 2025.

107. OpenAI, 'Operator System Card'. 2025: https://cdn.openai.com/operator_system_card.pdf.

108. Pariser, E., *The filter bubble: What the Internet is hiding from you*. 2011, London: Viking.

109. Persson, I., 'The Badness of Unjust Inequality'. *Theoria*, 2003. **69**: 109-124.

110. Pettit, P., *On the people's terms: A Republican theory and model of democracy*. 2012, Cambridge: Cambridge University Press.

111. Przeworski, A., 'Who Decides What Is Dem-

Electronic copy available at: knightcolumbia.org

ocratic?'. *Journal of Democracy*, 2024. **35**(3): 5-16.

112. Quinn, D.P. and J.T. Woolley, 'An Alternative View: Democracy Reflects the Preferences of Voters for Both Stability And Growth'. 2001.

113. Rau, E.G. and S. Stokes, 'Income Inequality and the Erosion of Democracy in the Twenty-First Century'. *Proceedings of the National Academy of Sciences*, 2025. **122**(1): e2422543121.

114. Rawls, J., *A theory of justice*. Rev. ed. 1999, Oxford: Oxford University Press.

115. Raz, J., *The morality of freedom*. 1986, Oxford: Clarendon Press.

116. Reid, M., et al., 'Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context'. *arXiv preprint arXiv:2403.05530*, 2024.

117. Ringel Morris, M., et al. 'Levels of AGI: Operationalizing Progress on the Path to AGI'. 2023. arXiv:2311.02462 DOI: 10.48550/arXiv.2311.02462.

118. Rini, R. and L. Cohen, 'Deepfakes, Deep Harms'. *Journal of Ethics and Social Philosophy*, 2022. **22**(2).

119. Rodrik, D., 'Why Does Globalization Fuel Populism? Economics, Culture, and the Rise of Right-wing Populism'. *National Bureau of Economic Research Working Paper Series*, 2020. **No. 27526**.

120. Ronfeldt, D., 'Cyberocracy is Coming'. *The Information Society*, 1992. **8**(4): 243-296.

121. Russell, S. and P. Norvig, *Artificial intelligence: A modern approach*. 3rd ed. 2016: Pearson Education.

122. Scheiring, G., et al., 'The Populist Backlash Against Globalization: A Meta-Analysis of the Causal Evidence'. *British Journal of Political Science*, 2024. **54**: 892-916.

123. Scheppele, K.L., 'Autocratic Legalism'. *The University of Chicago Law Review*, 2018. **85**(2): 545-584.

124. Schick, T., et al. 'Toolformer: Language Models Can Teach Themselves to Use Tools'. in *Advances in neural information processing systems*. 2023. Curran Associates, Inc.

125. Selbst, A.D. and S. Barocas, 'The Intuitive Appeal of Explainable Machines'. *Fordham Law Review*, 2018. **87**: 1085-1139.

126. Silver, D., et al., 'Reward Is Enough'. *Artificial Intelligence*, 2021. **299**: 103535.

127. Simon, F.M. and S. Altay, 'Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections'. *Knight First Amendment Institute*, 2025.

128. Socolow, M.J., 'Psyche and Society: Radio Advertising and Social Psychology in America, 1923-1936'. *Historical Journal of Film, Radio and Television*, 2004. **24**: 517-534.

129. Soest, C.V., 'Democracy Prevention: The International Collaboration of Authoritarian Regimes'. *European Journal of Political Research*, 2015. **54**: 623-638.

130. Stanley, M.L., et al., 'Resistance to Position Change, Motivated Reasoning, and Polarization'. *Political Behavior*, 2019. **42**: 891 - 913.

131. Stark, L., 'Algorithmic Psychometrics and the Scalable Subject'. *Social Studies of Science*, 2018. **48**(2): 204-231.

132. Sterling, T.D., 'Democracy in an Information Society'. *The Information Society*, 1986. **4**(1-2): 9-47.

133. Sumers, T., et al., 'Monitoring Computer Use via Hierarchical Summarization'. 2025.

Electronic copy available at: knightcolumbia.org

134. Sumers, T.R., et al., 'Cognitive Architectures for Language Agents'. *arXiv preprint*, 2023: https://arxiv.org/abs/2309.02427.

135. Summerfield, C., et al., 'How Will Advanced AI Systems Impact Democracy?'. *arXiv preprint arXiv:2409.06729*, 2024.

136. Susser, D., B. Roessler, and H. Nissenbaum, 'Online Manipulation: Hidden Influences in a Digital World'. *Georgetown Law Technology Review*, 2019. **4**: 1-45.

137. Susskind, J., *Future politics: Living together in a world transformed by tech*. 2018, Oxford: Oxford University Press.

138. Sutton, R.S. and A.G. Barto, *Reinforcement learning*. 2018, Cambridge, MA: MIT Press.

139. Tappin, B.M., et al., 'Quantifying the Potential Persuasive Returns to Political Microtargeting'. *Proceedings of the National Academy of Sciences*, 2023. **120**(25): e2216261120.

140. Temkin, L.S., *Inequality*. 1993, Oxford: Oxford University Press.

141. Tessler, M.H., et al., 'AI Can Help Humans Find Common Ground in Democratic Deliberation'. *Science*, 2024. **386**(6719).

142. The Heritage, F., *Mandate for leadership: The conservative promise 2025*. 2023, Washington, DC: The Heritage Foundation.

143. Trappey, C., 'A Meta-analysis of Consumer Choice and Subliminal Advertising'. *Psychology & Marketing*, 1996. **13**(5): 517-530.

144. Treisman, D., 'How Great Is the Current Danger to Democracy? Assessing the Risk With Historical Data'. *Comparative Political Studies*, 2023. **56**: 1924-1952.

145. Tsebelis, G., 'Veto Players and Law Production in Parliamentary Democracies: An Empirical Analysis'. *American Political Science Review*, 1999. **93**(3): 591-608.

146. Tufekci, Z., *Twitter and tear gas: The power and fragility of networked protest*. 2017, New Haven: Yale University Press.

147. Turkle, S., 'Confessional Machines', in *Life on the Screen: Identity in the Age of the Internet*. 1995, Simon & Schuster: New York.

148. Valmeekam, K., et al., 'Planning in Strawberry Fields: Evaluating and Improving the Planning and Scheduling Capabilities of LRM o1'. *arXiv preprint*, 2024: https://arxiv.org/abs/2410.02162.

149. Vaswani, A., et al., 'Attention Is All You Need'. *Advances in Neural Information Processing Systems*, 2017. **30**: https://arxiv.org/abs/1706.03762.

150. Viehoff, D., 'Democratic Equality and Political Authority'. *Philosophy & Public Affairs*, 2014. **42**(4): 337-375.

151. Voo, J., 'Driving Wedges: China's Disinformation Campaigns in the Asia-Pacific', in *Asia-Pacific Regional Security Assessment 2024: Key Developments and Trends*, S. International Institute for Strategic, Editor. 2024, Routledge for the International Institute for Strategic Studies: London. p. 79-97.

152. Wan, S., et al., 'CyberSecEval 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models'. *arXiv preprint*, 2024: https://arxiv.org/abs/2408.01605.

153. Wang, L., et al., 'A survey on large language model based autonomous agents'. *Frontiers of Computer Science*, 2024. **18**(6): 1-26.

154. Warren, M.E., 'A Problem-Based Approach

Electronic copy available at: knightcolumbia.org

to Democratic Theory'. *American Political Science Review*, 2017. **111**(1): 39-53.

155. Weizenbaum, J., *Computer power and human reason : from judgment to calculation*. 1976, San Francisco: W. H. Freeman.

156. Weizenbaum, J., 'ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine'. *Communications of the ACM*, 1966. **9**(1): 36-45.

157. Wiener, N., *The human use of human beings; Cybernetics and society*. 1950, Boston: Houghton Mifflin.

158. Williams, C., 'Public Psychology and the Cold War Brainwashing Scare'. *History & philosophy of psychology*, 2020. **21 1**: 21-30.

159. Winner, L., 'Mythinformation in the High-Tech Era'. *Bulletin of Science, Technology & Society*, 1984. **4**(6): 582-596.

160. Wright, N., 'How Artificial Intelligence Will Reshape the Global Order'. *Foreign Affairs*, 2018. **10**.

161. Xie, T. and I. Pentina, 'Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika'. 2022.

162. Xu, H., et al., 'Large Language Models for Cyber Security: A Systematic Literature Review'. *arXiv preprint arXiv:2405.04760*, 2024.

163. Yeung, K., '"Hypernudge": Big Data as Regulation by Design'. *Information, Communication & Society*, 2017. **20**(1): 118-136.

164. Young, I.M., *Inclusion and democracy*. 2000, Oxford: Oxford University Press.

165. Zhang, A.K., et al., 'Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models'. *arXiv preprint*, 2024: http://arxiv.org/abs/2408.08926.

166. Zittrain, J., 'Perfect Enforcement on Tomorrow's Internet', in *Regulating technologies: Legal futures, regulatory frames and technological fixes*, R. Brownsword and K. Yeung, Editors. 2008, Porland, OR: Hart Publishing. p. 125-156.

167. Zuboff, S., *The age of surveillance capitalism*. 2019, New York: Public Affairs.

168. Zuboff, S., 'You Are Now Remotely Controlled'. *The New York Times*. 2020.

Electronic copy available at: knightcolumbia.org

About the Authors

Seth Lazar is a professor of philosophy at the Australian National University, an Australian Research Council Future Fellow, and a Distinguished Research Fellow of the University of Oxford Institute for Ethics in AI. He has worked on the ethics of war, self-defense, and risk and now leads the Machine Intelligence and Normative Theory Lab, where he directs research projects on normative philosophy of computing. He was general co-chair for the Association for Computing Machinery (ACM) Fairness, Accountability, and Transparency conference 2022, and program co-chair for the ACM/Association for the Advancement of Artificial Intelligence's AI, Ethics, and Society conference in 2021 and is one of the authors of a study by the U.S. National Academies of Science, Engineering, and Medicine on the ethics and governance of responsible computing research. He gave the 2022 Mala and Solomon Kamm lecture in ethics at Harvard University and the 2023 Tanner Lectures on AI and human values at Stanford University.

Mariano-Florentino (Tino) Cuéllar—a law professor and public servant with broad experience in international and domestic policy, the justice system, education, and philanthropy—is the president of the Carnegie Endowment for International Peace. A scholar of transnational regulatory and security problems, American institutions, and technology's impact on law and government, he previously served as a justice on the Supreme Court of California, the highest court of America's largest judiciary. He is the first Mexican immigrant ever to serve in this capacity. Previously, he was the Stanley Morrison Professor at Stanford Law School and director of Stanford University's Freeman Spogli Institute for International Studies. He served two U.S. presidents in a variety of roles in the federal government, including as special assistant to the president for justice and regulatory policy in the Obama administration.

Electronic copy available at: knightcolumbia.org

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, policy advocacy, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Electronic copy available at: knightcolumbia.org