KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

Anticipatory AI Ethics

Steering AI ethics towards the technological horizon

By Seth Lazar

April 24, 2025

Sébastien A. Krier using Midjourney 6.1

Abstract

Anticipating and steering the societal impacts of frontier artificial intelligence (AI) systems demands a kind of speculation that often faces both moral and epistemic objections. For example, some contend that "anticipatory AI ethics" amplifies AI hype and draws attention away from present-day harms or that it rests on dubious technological determinism and shaky, predictably false predictions. In response, this essay defends an approach to anticipatory ethics grounded in epistemic humility and a clearly defined technological horizon: the

range of possible worlds that can be reasonably well-understood given what we know now about AI and the social structures into which it will be deployed. It then raises a crucial question: Are genuinely transformative AI systems within or beyond this horizon?

Introduction

Artificial intelligence (AI) is driving rapid societal transformation. Recent experience has starkly demonstrated that merely reacting to technological change—whether from AI or from digital innovation more broadly—is insufficient. We urgently need to proactively forestall rather than just respond to technological impacts.[1] Yet this pre-emptive approach poses genuine challenges. This essay describes and defends a particular conception of "anticipatory ethics" (Johnson, 2011), which seeks to identify ethical issues raised by emerging technologies ex ante, using that insight to actively shape those technologies for the better.

I will first consider the main objections to anticipatory ethics in the context of AI. I will then present an approach that should avoid the objections, at least when applied within what I will describe as the "technological horizon"—roughly, the set of possible worlds we can reliably envision based on projecting forward from our current technological capabilities and institutional frameworks. I will close by considering whether this opens or closes the door to applying methods of anticipatory ethics to radically transformative AI.

What Is Wrong with Anticipatory Ethics?

Looking to technology's future has a long history that includes other branching paths—such as "speculative ethics."[2] As long as ethicists have had their eyes on what is next, they have been upbraided with similar shortcomings.[3] Some objections are essentially moral arguments against engaging in anticipatory ethics. Other objections are epistemic.

The moral arguments against anticipatory ethics are sometimes vituperative. For example, anticipating the societal impacts of powerful new AI systems involves conditioning on the claim that they are both powerful now and likely to soon become even more so. This means expressing at least some agreement on AI capabilities with the people most substantially materially invested in convincing the world about AI's radical promise. Doing so presents a danger—indeed, according to some, constitutes a realized wrong—of lending credence to irresponsible marketing hype.[4] Nor is this specific to AI—in the 2000s, ethicists of technology raised a similar complaint against anticipatory ethics as applied not to computing but to nanotechnology.[5]

Perhaps because they are alive to this moral critique of their research practices, some scholars engaged in anticipatory ethics overcompensate. How could they be making common cause with the AI boosters if their forecasts are apocalyptic? Hyperbolic critique of new technologies is not confined to anticipatory ethics. Criticism of current technologies

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

often strays into what Vinsel (2021) calls "CritiHype."[6] But, at least for AI, there has been a recent tendency for anticipatory ethics to become associated with narrow attention on the worst possible outcomes to which AI advances might lead.[7] This narrow focus grounds two substantive criticisms. First, anticipatory ethicists are focusing on hypothetical future downsides at the expense of attending to current harms (again, the same things were said during the last major anticipatory ethics debates that focused on nanotechnology two decades ago).[8] Second, by focusing only on the worst possible outcomes from AI, ethicists are failing to offer any guidance for the much more likely worlds in which those extreme outcomes are avoided, but society is still irreversibly changed (Lazar and Nelson, 2023; for a similar concern in the bioethics case, see Carter et al., 2009).

Moral critiques of competing research agendas sometimes feel ideological and overwrought (pluralists are disinclined to tell others what to research or talk about). But there are more tempered epistemic objections to anticipatory ethics as well. The most compelling highlights its tendency to slip into technological determinism, especially the view that one can reliably predict the all-things-considered outcomes of some new technology just by considering its internal properties (Mumford, 1964; Winner, 1980). This disregards not only the many different sociotechnical systems with which any newly developed technology must interact but also the varied and unpredictable ways in which humans inevitably repurpose and redirect new technologies.[9]

Much of the theory of existential risk (x-risk) from AI—a canonical form of anticipatory ethics—is explicitly deterministic in this sense. For some in this camp, we do not need to know anything about the world into which superintelligent AI will be deployed to know that, absent some profound discovery between now and then, it will end human life on Earth. At a more prosaic level, but still in a similar way, the advent of capable large language models (LLMs) led some to prematurely project that synthetic content would singlehandedly destroy democratic institutions (Kreps and Kriner, 2023; for a sober analysis of generative AI's impact on 2024's "year of democracy," see Simon et al., 2024). Going further back, early predictions of the impact of the internet and algorithmic recommendation on political discourse imagined a world in which just because technology could conceivably trap us in echo chambers and filter bubbles, it would unavoidably do so (Sunstein, 2001; Pariser, 2011). Things have turned out to be much more complicated than that (Bruns, 2019). Similarly, alarmed by the disturbing potential of micro-targeted advertising and algorithmic recommendation, Zuboff (2020) fantasized that we were all to become marionettes on strings, in the control of tech Geppettos. Again, this gave too much credence to the claims of those selling ads and not enough to the research showing that microtargeted advertising has modest, if any, effects (Hwang, 2020; Benn and Lazar, 2021).

Of course, the most obvious epistemic shortcoming of anticipatory ethics is simply that it takes the future as its object. It is hard to say actionable, well-grounded, falsifiable things about the future outside of either very broad macro trends or else narrowly quantifiable domains (and even in those, success is limited). It is harder still to find a futurist who holds

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

himself accountable for his past predictions; ironically, futurism rarely stands the test of time. Our ability to predict complex social phenomena is, on the whole, not great.

The Necessity of Anticipatory Ethics for AI

In the face of these limitations, why not just abandon anticipatory ethics or else leave it to the bloggers? The answer is obvious. If new technologies are likely to cause significant societal harm before we can develop adequate post hoc measures to remedy those harms, then we need to design those technologies to mitigate those risks. Anticipatory ethics guides that process. We cannot really escape it. The real question, then, is can we do it better?

Examples abound, but consider one instructive case. Today, it is widely recognized that carelessly applying machine learning (ML) to decision-making risks discriminating against disadvantaged minorities. Yet, the scale of actual harms resulting from discriminatory ML algorithms has so far been less severe than someone familiar with the literature on discriminatory AI might reasonably have feared.[10] These two claims might initially appear in tension: One could mistakenly conclude that concerns about discriminatory AI were overblown, given the comparatively modest realized societal harms. But this would overlook the crucial role played by researchers engaging in anticipatory ethics, who demonstrated clearly and proactively that, if deployed without care, ML decision-making models would exhibit predictable biases. By explicitly exposing these risks, proposing effective mitigation methods, and advocating against deploying irreparably biased systems, anticipatory ethics scholars meaningfully reduced the scale of potential harms. In short, discriminatory impacts from ML have been reined in precisely because AI ethics researchers anticipated these harms and intervened ex ante (or at least early).[11]

Anticipatory ethics is especially indicated in problem domains that have these features: (1) Rapid technological progress is underway; (2) The gap between fundamental research discoveries and society-wide deployment could be small; (3) The variance between the possible outcomes (how bad or how good they can get) is high; and (4) There are levers with which we could influence those outcomes ex ante.

The current state of AI clearly meets condition (1). In mid-2022, it was common to describe progress in AI as a patchwork of areas of extremely narrow and brittle superhuman competence. A model could learn to beat any human at chess or go but would not be able to pick up a new game without first playing itself millions of times, and if you just change the board or the rules subtly, it would entirely fail to adapt. Many thought true linguistic competence—or expertise in the visual arts, music, the spoken word, or broad common sense and ethical judgment—would be artificial general intelligence (AGI)-complete (that is, would require AGI to be robustly performed).[12] Yet, we now have incredibly capable general-purpose models that can not only generate almost any content a talented human could generate but are progressively more capable of acting in practically rational ways in arbitrary domains. We are on a path towards AI with human-like breadth in competence,

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

paired with discrete capabilities that are, in many respects, very clearly superhuman. At this time, it is hard to confidently place *any* upper bound on the level of capabilities that will be realized within the present paradigm.

Prior to the advent of reasoning models, and the associated use of reinforcement learning to teach LLMs how to use additional tokens at inference time to resolve hard problems, I think one could plausibly argue that autoregressive LLMs, in principle, had an upper bound to their potential, due at least to two facts: their inability to allocate compute differentially to parts of their response conditioned on the importance of that part of the response; and their path-dependent nature. An LLM applies the same amount of compute to every token that it generates: If it faces a branching path where one token leads it to one conclusion, another to the opposite, it does not "think" about that token more than it "thinks" about whether to use a definite or indefinite article. But rational thought does require one to allocate computation in proportion to the difficulty or importance of a problem. This weakness of LLMs is compounded by their autoregressive nature, which means that they always condition on the last token that they generate, so they are not, by default, prone to reevaluate a line of reasoning if it leads to a dead end. Instead, they just make the best of it, given that they have to condition on it. Reasoning models that can spend more tokens at test time are, despite still being autoregressive LLMs, much less constrained in these ways. They can backtrack and revisit a key branching moment in their reasoning and choose differently. This means both that they are not so path-dependent and that they can allocate computation more closely in proportion to a problem's difficulty and importance. It does not mean we should expect this method to provide unbounded, undiminishing returns. But it does substantially mitigate two in-principle obstacles to fundamental progress.

Condition 2 is contentious. As Narayanan and Kapoor (2025) note, frontier research and development of any given technology often takes time to translate into realized products and services. If that is true with AI, then we potentially have more room for a "wait and see" posture. However, general economic inertia aside, current AI systems admit of a much narrower gap between discovery and deployment than most previous technologies. Every sector of society depends on computational systems owned or controlled by companies that are either conducting or at least financing frontier AI research. New AI models are structurally compatible with many of these software systems. Deployment to billions of people all around the world requires no more than an over-the-air update. See, for example, Microsoft's deployment of new OpenAI models in CoPilot, Amazon's upgrade of Alexa to Alexa+ using Anthropic's Claude model (as well as their own Nova models), or indeed OpenAI's own 400,000,000 monthly active users (and growing) (Washenko, 2025; Lee, 2025; Reuters, 2025). Or go back further to the early days of digital platform companies—while Page and Brin's PhD research on PageRank took around five years to reach hundreds of millions of users, by 2010, breakthroughs in research to predict click-through rates could go from presentation at ICML to deployment to 700,000,000 Facebook users within just months (Hao, 2021). Meanwhile, an army of start-ups has built LLM-based software

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

predicated on the underlying models achieving a particular level of performance in the future. When sufficiently capable and efficient models are developed, there will be vessels waiting to carry them directly to market across many different industries. Undoubtedly, some forms of diffusion will be laggy, but there is likely to be a "jagged frontier" in which, in some areas, very powerful AI systems might be deployed overnight. This gives reason to favor more anticipatory approaches.

Condition 3 is clearly met. While the concerns of AI x-risk catastrophists might seem overblown to some, it is indeed hard to overstate the scale of the possible societal impacts of AI systems, even clearly within the current technological horizon. Even relatively prosaic AI could put us on a path to increase the rate of global growth by an order of magnitude (Erdil and Besiroglu, 2023). It could induce radical transformations in how (and whether) we work (Susskind, 2020). It could decisively change the balance between cyberattack and defense, rendering networked systems intrinsically insecure (Guven, 2024).

On a less prosaic level, from the first recorded civilizations to the present day, humankind has always aspired to invest dumb matter with spirit and agency—from the golems of Jewish lore to the alchemists' homunculi, from Frankenstein's monster to the original animating impulse of the invention of computing: to create a machine that could think (Turing, 1950; Goodman, 2023). In the 5,000 years of recorded human history, we happen to be alive at the moment when at least one realization of that deeply ingrained human dream is, for better or worse, actually potentially within reach—even according to many skeptics.[13] The stakes are high.

Lastly, there are indeed levers that we can pull to secure better outcomes, perhaps uniquely so. Consider, by contrast, taking the anticipatory ethics approach to nanotechnologies, human cloning, or medicine. As ethicists, we have a pretty nominal ability to actually shape the outcomes of those endeavors for the better. We can inform judgments about whether to permit the scientists to proceed or about the costs and benefits of some prospective intervention. However, we typically cannot intervene directly in the process of cloning, or calibrate the side effects of the new treatment such that they better serve human values. By contrast, the design of advanced AI does give us explicit opportunities to shape its societal impacts by shaping AI systems themselves.[14] For example, those who release the first really capable AI agents will use various design measures to ensure that, when deployed, they are at least to some degree secure, safe, and trustworthy (or some similar set of evaluative terms). How they use those measures is determined by their expectations of the societal impacts their agents might have. Anticipatory ethics is not only a cudgel with which to hold the companies careering into the future to account; it can also provide a roadmap or blueprint to avoid the worst consequences that it has foretold.

Importantly, this is not just about ethicists being deeply involved in the design of more capable AI systems to ensure, for example, that they are aligned with human values. We are now building highly capable general-purpose AI systems. They can, in principle, be put to any use. But they will be used only for purposes that we have the imagination to

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

task them with. A crucial role for anticipatory ethics, now, is to influence what kinds of products and infrastructure we build with and for powerful AI systems (Lazar et al., 2024; Chan et al., 2025; Kapoor et al., 2025). This is an uncomfortable position for especially academic researchers to be in—our default mode is critique and identification of risks. Even more speculative anticipatory ethics will often receive a pass through peer review when it confines itself to calling out the downsides of some new technology. It would be unusual to say, "Unless you build it, and it indeed has those downsides, your claims lack epistemic warrant." The same is not true for work that invokes the moral imagination to conceive of new uses for highly capable AI systems. We should shake this bias.

Anticipatory Ethics Within the Technological Horizon

Progress in AI seems to demand anticipatory AI ethics. But, as argued above, this involves both moral and epistemic risks: fanning hype, being hyperbolically critical, spreading too thinly and misguidedly the zero-sum resource of attention, spurious technological determinism, and bad predictions. Let us tackle the epistemic risks first before checking whether epistemically responsible anticipatory ethics avoids moral objections, too.

Step 1: Anticipatory ethics need not traffic in *unconditional all-things-considered predictions*. These are what you get from pointy-hatted ladies rhyming around a cauldron. Philosophers of AI without the gift of the eye are as unlikely as other mortals to divine comprehensive pictures of how the world will go, all things considered, under AI (those who have the eye should get out of philosophy and into finance). While there are ways to draw on foresight studies, prediction markets, and other resources to try to capture some of the Weird Sisters' magic, we can also usefully adopt a more epistemically conservative approach—especially if we stay within the technological horizon.[15]

The technological horizon for AI marks the boundary of possible worlds that we can reasonably understand based on what we know now about AI systems and about our current social, political, and economic structures. Both of these constraints are important. If we do not hold relatively fixed the environment into which these systems will be deployed, then we radically increase our uncertainty: Instead of just having to forecast AI capabilities, we have to conjure up a whole ecosystem and how AI would impact *that*. Equally, if we just imagine purely hypothetical future AI systems, then the spectrum of possibilities explodes with few rules to constrain our speculation. Still, more importantly, we have no levers to intervene in purely hypothetical systems (short of shutting down all AI research). If we focus instead on systems that are plausible extensions of those we can work with today, then we have realistic research pathways by which to intervene and steer those systems.

With this in place, we can then engage in *constrained* analysis: instead of asking, "What will the impacts of these systems be?" we identify particular features of those systems that, given deployment in that environment, are likely to increase the probability of negative or positive outcomes. These discrete *hazards* and *opportunities* could be described, in terms of Science and Technology Studies, as *affordances*—properties of the system that

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

make certain outcomes more or less likely without necessitating them (Davis, 2020).

Take language model agents (LMAs) for example.[16] These are AI systems capable of undertaking complex tasks without direct human supervision, in which LLMs provide the essential cognitive resources. There are many possible futures for LMAs. They might stall out for years, unable to surmount the "capability-reliability gap" (that is, while it might be possible to create impressive demos of LMAs performing complex tasks, we might find that getting production-level LMAs to reliably perform those tasks in practice is much harder).[17] Or we might soon wind up with widespread LMAs that can effectively use a computer to perform more or less all the same tasks that a competent human user could perform. Or LMAs might blow past the human baseline in every dimension. These are all open futures, and one can make a case for almost any distribution of probabilities among them; nor are they jointly exhaustive of the possibilities. The interesting question is not just which is more likely, but what would the societal impacts of each of these scenarios be. Given that all are based on plausible extensions of our existing technologies—they are within the technological horizon—we have plenty to go on in making these projections.

Of course, understanding the technology alone is not enough. We must also adopt a reasonable model of the world into which that system will be deployed. This approach allows us also to attend to the broader structural forces that will shape (and in turn be shaped by) its deployment and adoption. The societal impacts of any technology are a function not just of the tech itself but of the environment into which it is injected (including the people within it) (Winner, 1980). Existing regulatory frameworks, political alignments, economic incentives, cultural norms, and practices all contribute. Anticipatory ethics should individuate the features of some assumed technological advance—the properties of that technology that are likely to be causally effective—and identify those that are, given the environment into which the system is deployed, likely to be more societally beneficial and those that will be more harmful. The goal is not to make an all-things-considered judgment but to highlight discrete hazards and opportunities that can be mitigated or exploited when designing and deploying these systems. This obviously relies on pairing anticipatory ethics with anticipatory social science (Nelson and Banks, 2018).

Additionally, because we are simply identifying hazards and opportunities of the technology, we can avoid the fallacy that its nature unilaterally dictates the societal outcomes to which it will lead, with no role therein for human agency or other factors. In fact, far from accepting technological determinism, anticipatory ethics is precisely aimed at foregrounding human agency by identifying key points for intervention (Johnson, 2011).

So, anticipatory ethics should take as follows its basic question: Given our current institutional context, what hazards and opportunities would arise from capabilities that plausible extensions of today's AI systems might realistically acquire?

For example, LMAs will be deployed into an economy dominated by large technology firms, many of them platform companies that exercise considerable power over the users

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

of those platforms (Kapoor et al., 2025). Since LMAs will radically enhance and extend the agency of their principals—insofar as LMAs will be able to undertake complex sequences of tasks without direct human supervision—we should expect the deployment of AI agents in the platform economy to create strong centralizing tendencies, which threaten to further concentrate power. However, this is just one feature of LMAs and the environment into which they will be deployed; it is not a summary judgment that we are inevitably headed for a world of platform agents.

But what about the moral objections to anticipatory AI ethics? In light of the foregoing, they indeed seem overwrought. Conditional projections of the consequences of a particular set of capabilities being realized cannot plausibly contribute to AI hype. Nor can just yelling "hype" serve as a counterargument to the demonstrated progress that has been made in AI over the last two years. Indeed, perhaps we just happen to be alive at a time when accurately assessing technological capabilities involves language that would otherwise be thought hyperbolic. We should certainly not describe states of affairs inaccurately to avoid the appearance of hyperbole. The complaint that attention is a zero-sum resource, and that focusing on future risks from AI systems distracts from current harms, is wrong on every level. The allocation and attraction of attention is a function of political aptitude, among other things. Whether advocating for one issue detracts from or complements advocating for another is an empirical matter unlikely to be resolved by summary a priori judgment. Moreover, future direct risks from AI systems might, in fact, be morally more serious than those from current systems—dismissing that possibility out of hand is irrational; anticipatory ethics is our means for deciding whether it is true. Nor should sensitivity to the full range of AI risks prevent us from considering what happens if AI does not kill us all—this is simply a bad choice of focus; we can choose differently. Indeed, if we do not pay sufficient attention to managing the societal impacts of AI short of existential threats, then if and when future AI systems do threaten the survival of humanity, extinction might come as a blessed relief for those of us still left.

Where Is the Technological Horizon?

Researchers exploring AI's societal impacts need to look towards the horizon and not only focus on "current harms." They can do so in an epistemically responsible way. But are some of the prospective risks from AI systems, in fact, beyond the technological horizon and, as such, much harder objects for anticipatory ethics? My running example throughout has been LMAs, AI agents that are clearly within the technological horizon. Let us now define "transformative AI" (TAI) as transformative in two senses: It causes rapid, radical change in more-or-less every sector of society; and it has unprecedented capabilities whose contours (and effects) are hard to anticipate in detail from our present epistemic standpoint. Can anticipatory ethics be applied to TAI?

To illustrate: AI agents that are truly autonomous, able to reliably undertake not only tasks but whole roles, and perhaps even self-actuating and capable of setting their own goals

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

would count as TAI. Can we do respectable anticipatory ethics conditioned on the possibility of developing rationally autonomous AI agents that are more capable than humans in every dimension where comparison is possible? How can we sensibly predict just what capabilities will be enhanced, and what will be the consequences of sharing the world with autonomous entities vastly more intelligent than us (in some opaque sense)?

Similarly, some think that AI will soon radically accelerate scientific research, providing what Dario Amodei (2024) calls "a 'country of geniuses in a datacenter.'" Can this kind of TAI be subjected to anticipatory ethics? Amodei makes a heroic attempt, but should we not be skeptical of any attempt to forecast just what an AI system that enables radical scientific progress could enable us to do? It would be like trying to predict the outcome of asking a genie for infinite wishes—anticipating at the outset what the world would be like under unconstrained scientific progress seems impossible.

It is tempting to think this is a bridge too far. When anticipatory ethics becomes too speculative, its rules of engagement become opaque. Beyond compliance with basic logic and the laws of nature, perhaps there are none. One might think this a kind of scholasticism: diverting but not really relevant for guiding action since (a) it is unlikely to recommend specific actions and (b) the dominant strategy is likely to be acquiring more robust evidence before deciding (Casper et al., 2025). While earlier speculation might be vindicated by that new evidence, it still proves largely redundant once we have better epistemic resources for our decisions.

Anticipatory ethics that looks over the technological horizon is probably, in general, a bad strategy.[18] The thing is, I suspect that TAI is not, in fact, beyond the technological horizon. Or at least, this is a reasonable topic for debate.

Can we today provide resources that will be useful after significant AI progress has been made to help society respond to and shape the impacts of TAI? Do we have enough information about the world into which TAI will be deployed, and about the likely contours of TAI, that we can create the philosophical equivalent of standard operating procedures (SOPs) that can be adapted and then usefully put into practice on short notice when they are needed? Do we have reason to think that these resources will actually prove useful in the event of that societal transition taking place, rather than just being heuristics that we discard once we have more concrete information?

I suggest that we should answer all these questions affirmatively because there is a non-trivial prospect of truly transformative AI being developed in the next decade.[19] If so, TAI will likely be architecturally substantially similar to existing AI systems (in at least the same way that, say, DeepSeek's R1 model is architecturally similar to GPT-2). So, I think our understanding of the current technology does give us resources with which to condition our ethical work on a wide range of plausible sets of capabilities that future AI systems might realize. Because the world and its social structures are—for all the turmoil of 2025 so far—comparatively stable, at least at the macro-level, we broadly know enough

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

about the environment into which TAI will be deployed to make sensible projections about what the societal impacts will be, conditional on a particular level of capability being reached. Most importantly for the project of anticipatory ethics, because TAI will most likely be developed by the companies currently at the frontier of AI research, we actually have means to influence its development now. We can also build up societal resilience and create effective SOPs that we can customize based on new information as we acquire it (Bernardi et al., 2024).

But even if TAI is sufficiently within the technological horizon to be a proper object for anticipatory ethics, that does not elide the fact that it is obviously harder to pin down than the logical next step in AI capabilities. What, then, does good anticipatory ethics about TAI look like?

A methodologically conservative approach is to anchor our speculation about society under TAI in some other individual's or organization's "technological imaginary." By engaging in internal critique of that imaginary, we can at least shed light on the action-relevant question of whether we should tolerate or support the aspiration to realize that aim now.

For example, any AI lab that claims to be aiming to build "safe" or "aligned" artificial superintelligence (ASI) but which has not considered how their prospective ASI could be subjected to meaningful and enduring democratic control is putting our democratic freedoms at risk (Lazar and Pascal, 2024). This is because an ASI could quite easily be safe or aligned and still be decisively detrimental to democracy insofar as it relieves democratic publics of control over their individual and collective lives. For each task or domain of our lives that we automate, we risk losing something in the process. By relying on an ASI to resolve our problems, we risk becoming incapable of addressing them ourselves (Bainbridge, 1983; Vallor, 2015). Perhaps it is possible to design an ASI that preserves democratic self-rule. However, it will not happen by accident. So unless the frontier AI companies are explicitly thinking about how their prospective future ASIs will be subjected to meaningful democratic control, democracy is in trouble.

But this kind of internal critique can get us only so far. A less conservative approach takes seriously the challenge of forecasting the capabilities that TAI might have, and the social and other relations it might entail. Instead of attempting to predict how these might play out in practice, it shows how these capabilities and impacts could potentially undermine central presuppositions of our existing institutions as well as our social and political theories. The goal is not so much to forecast what society under TAI will be like but to give a spec sheet for the task of political and other philosophers when we do know more about what it will be by demonstrating the likely inadequacy of our existing theories for that task.

For example, a world with TAI will likely involve enabling either (a) the radical alteration of human capacities and limitations or (b) the creation of non-human agents that plausibly satisfy pre-existing criteria for moral personhood (or both). Our existing moral and

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

political theories, as well as many of our arguments for them, are conditioned on the basic (rough) descriptive equality of all moral persons, as well as our mutual vulnerability and mutual dependence. These empirical assumptions are all likely to be undermined by TAI. If we create AI agents that are moral persons, then they will have radically different capabilities, limitations, and identity conditions than do human agents (e.g., software agents would be able to fork, merge, multiply, and divide themselves in ways that are not feasible for humans). The same is true if we create TAI systems that are not themselves moral persons but radically change the capabilities, limitations, and perhaps even identity conditions of at least some people who use them.

Similarly, we can note that existing institutions and conventions have evolved and been developed for the purposes of facilitating interactions among humans with our characteristic capabilities, limitations, and identity conditions. In a world where AI agents interact with orders of magnitude more rapidly, frequently, and consequentially than human agents do, new institutions and conventions—from money to markets to language—will emerge to facilitate those interactions. We, therefore, need to be ready for the contingency of social facts (that is, the social phenomena that we construct through convention and implicit collective assent) to be acutely and suddenly exposed and for new social facts to be created (Searle, 1997).

A third, more ambitious approach attempts to conduct first-order anticipatory ethics for a world with TAI and perhaps even to build a moral, social, or political theory directly tailored to that world. This is especially hard to do, not only because the properties of TAI are hard to predict but because TAI, of its nature, will impact every sector of society. It is, therefore, hard to abide by the methodological restriction of avoiding all-things-considered predictions.

Still, there remains an important distinction between simply imagining what society might look like a decade (say) after TAI's arrival and explicitly constructing a transition function for how TAI could drive structural changes from our current social fabric to radically new forms. Even if it is still inherently epistemically risky, the latter is more justified than the former. Moreover, if we condition on the assumption that TAI will affect everything, then the greater epistemic risk would lie in attempting the more discrete approach that holds as much as possible constant while just varying AI capabilities.

This observation poses an interesting challenge for contemporary moral, social, and political theory, given that it is long since out of the habit of forming synoptic, integrated grand visions—indeed, it has spent decades excoriating this kind of agenda. Skepticism of grand narratives and a propensity for conservative miniaturism might be forgivable in more stable times. An age of genuine revolutions may demand greater ambition.

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

Conclusion

This essay advocates for anticipatory ethics in general and for its application to AI in particular. Indeed, I think that rapid progress in the development of, for example, AI agents is evidence not only of the necessity but also the tractability of anticipatory ethics. Those concerned with the societal impacts of AI must at least look to the horizon and not focus only on current harms from deployed systems.

Anticipatory ethics of transformative AI is perhaps harder to epistemically justify. The previous section ultimately poses a litmus test that will divide those who favor applying anticipatory ethics to TAI from those who do not. Should we expect TAI to arise from systems that are relevantly similar to those we work with today? Should we expect the rest of the world to look more or less as it does today when TAI is developed? Affirmative answers to these questions vindicate the project of anticipatory ethics for TAI. Though undoubtedly more challenging than focusing on nearer-term, more discrete AI developments, it is both feasible and urgent. If one instead answers those questions negatively, then anticipatory ethics for TAI is neither feasible nor urgent. In either case, the real question is not whether we should be engaged in anticipatory ethics at all but whether TAI's advent is within our technological horizon. While reasonable experts clearly disagree on this point, I think that it is. As such, there is much work to do.

References

Amodei, D. 2024. "Machines of Loving Grace."

Bainbridge, L. 1983. "Ironies of Automation." Automatica 19 (6): 775-79.

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

Bernardi, J., G. Mukobi, H. Greaves, L. Heim, and M. Anderljung. 2024. "Societal Adaptation to Advanced AI." arXiv preprint: https://arxiv.org/abs/2405.10295.

Bostrom, N. 2007. "Technological Revolutions: Ethics and Policy in the Dark," in Nigel M. de S. Cameron and M. Ellen Mitchell (eds.), Nanoscale: Issues and Perspectives for the Nano Century. Hoboken, NJ: John Wiley & Sons, 129-52.

Brey, P. 2012. "Anticipatory Ethics for Emerging Technologies." NanoEthics 6 (1): 1-13.

Bruns, A. 2019. Are Filter Bubbles Real? Cambridge: Polity Press.

Buolamwini, J. and T. Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in A. Friedler Sorelle and Wilson Christo (eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency (81; Proceedings of Machine Learning Research: PMLR), 77-91.

Burgess, J. 2023. "Everyday Data Cultures: Beyond Big Critique and the Technological Sub-

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

lime." AI & Society 38 (3): 1243-44.

Carter, A., P. Bartlett, and W. Hall. 2009. "Scare-Mongering and the Anticipatory Ethics of Experimental Technologies." The American Journal of Bioethics 9 (5): 47-48.

Casper, S., D. Krueger, and D. Hadfield-Menell. 2025. "Pitfalls of Evidence-Based AI Policy." arXiv preprint: https://arxiv.org/abs/2502.09618.

Chan, A., K. Wei, S. Huang, N. Rajkumar, E. Perrier, S. Lazar, G. K. Hadfield, and M. Anderljung. 2025. "Infrastructure for AI Agents." arXiv preprint arXiv:2501.10114.

Chan, A., R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj. 2023. "Harms from Increasingly Agentic Algorithmic Systems," Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL: Association for Computing Machinery), 651–66.

Collingridge, D. 1980. The Social Control of Technology. New York: St. Martin's Press.

Davis, J. L. 2020. How Artifacts Afford: The Power and Politics of Everyday Things. Cambridge, MA: The MIT Press.

Erdil, E. and T. Besiroglu. 2023. "Explosive Growth from AI Automation: A Review of the Arguments." arXiv preprint: https://arxiv.org/abs/2309.11690.

Floridi, L. and A. Strait. 2020. "Ethical Foresight Analysis: What It Is and Why It Is Needed?" Minds and Machines 30: 77-97.

Gabriel, I., A. Manzini, G. Keeling, L. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, S. El-Sayed, S. Brown, C. Akbulut, A. Trask, E. Hughes, A. Bergman, R. Shelby, N. Marchal, C. Griffin, J. Mateos-Garcia, L. Weidinger, W. Street, B. Lange, A. Ingerman, A. Lentz, R. Enger, A. Barakat, V. Krakovna, J. Siy, Z. Kurth-Nelson, A. Mccroskery, V. Bolina, H. Law, M. Shanahan, L. Alberts, B. Balle, S. De Haas, Y. Ibitoye, A. Dafoe, B. Goldberg, S. Krier, A. Reese, S. Witherspoon, W. Hawkins, M. Rauh, D. Wallace, M. Franklin, J. Goldstein, J. Lehman, M. Klenk, S. Vallor, C. Biles, M. Morris, H. King, B. Agüera Y Arcas, W. Isaac, and J. Manyika. 2024. "The Ethics of Advanced AI Assistants." Google DeepMind. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf.

Goodman, L. 2023. "Alchemy, Mythology, and Artificial Intelligence What Has Enchanted—and Alarmed—Imagination." Renovatio: The Journal of Zaytuna College.

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

Hao, K. 2021. "How Facebook Got Addicted to Spreading Misinformation." MIT Technology Review, Mar 11.

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

Helfrich, G. 2024. “The Harms of Terminology: Why We Should Reject So-Called ‘Frontier AI’.” AI and Ethics 4 (3): 699-705.

Hendrycks, D., M. Mazeika, and T. Woodside. 2023. “An Overview of Catastrophic AI Risks.” arXiv preprint: https://arxiv.org/abs/2306.12001.

Hwang, T. 2020. Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet. New York: Farrar, Straus and Giroux.

Johnson, D. G. 2011. “Software Agents, Anticipatory Ethics, and Accountability.” The Growing Gap between Emerging Technologies and Legal-Ethical Oversight, Springer Netherlands: 61-76.

Kapoor, S., N. Kolt, and S. Lazar. 2025. “Position: Resist Platform-Controlled AI Agents and Champion User-Centric Agent Advocates.” arXiv preprint.

Kapoor, S., B. Stroebl, Z. S. Siegel, N. Nadgir, and A. Narayanan. 2024. “AI Agents That Matter.” arXiv preprint: https://arxiv.org/abs/2407.01502.

Kolt, N. 2024. “Governing AI Agents.” SSRN: https://dx.doi.org/10.2139/ssrn.4772956.

Kreps, S. and D. Kriner. 2023. “How AI Threatens Democracy.” Journal of Democracy 34 (4): 122-31.

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

Lazar, S. and A. Nelson. 2023. “AI Safety on Whose Terms?” Science Magazine 381 (6654): 138-38.

Lazar, S. and A. Pascal. 2024. “Agi and Democracy.” Allen Lab for Democracy Renovation.

Lazar, S., L. Thorburn, T. Jin, and L. Belli. 2024. “The Moral Case for Using Language Model Agents for Recommendation.” arXiv preprint: https://arxiv.org/abs/2410.12123.

Lee, D. 2025. “Amazon’s new Alexa is like an easier-to-use ChatGPT. It could be huge,” Australian Financial Review, Feb 27. https://www.afr.com/technology/if-the-new-alexa-works-as-advertised-amazon-s-made-it-relevant-again-20250227-p5lfjd.

Lucivero, F., T. Swierstra, and M. Boenink. 2011. “Assessing Expectations: Towards a Toolbox for an Ethics of Emerging Technologies.” NanoEthics 5: 129-41.

Mumford, L. 1964. “Authoritarian and Democratic Technics.” Technology and Culture 5 (1): 1-8.

Narayanan, A. and S. Kapoor. 2025. “AI as a Normal Technology.” Knight First Amendment Institute.

Nelson, A. and D. Banks. 2018. “Concept Note Detailing Anticipatory Social Research.” Social Science Research Council.

Noble, S. U. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. New

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

Susskind, D. 2020. A World without Work: Technology, Automation, and How We Should Respond. New York, NY: Metropolitan Books/Henry Holt & Company.

Swoboda, T., R. Uuk, L. Lauwaert, A. P. Rebera, A.-K. Oimann, B. Chomanski, and C. Prunkl. 2025. "Examining Popular Arguments against AI Existential Risk: A Philosophical Analysis." arXiv preprint: https://arxiv.org/abs/2501.04064.

Turing, A. M. 1950. "Computing Machinery and Intelligence." Mind 59: 433-60.

Vallor, S. 2015. "Moral Deskilling and Upskilling in a New Machine Age." Philosophy & Technology 28 (1): 107-24.

Vallor, S. 2024. The AI Mirror: How to Reclaim Our Humanity in the Age of Machine Thinking. New York, NY: Oxford University Press.

Varoquaux, G., A. S. Luccioni, and M. Whittaker. 2024. "Hype, Sustainability, and the Price of the Bigger-Is-Better Paradigm in AI." arXiv preprint: https://arxiv.org/abs/2409.14160.

Vinsel, L. 2021. "You're Doing It Wrong: Notes on Criticism and Technology Hype," Medium.

Washenko, A. 2025. "Microsoft Copilot offers Voice and o1-powered Think Deeper for free," Engadget, Feb 25. https://www.engadget.com/ai/microsoft-copilot-offers-voice-and-o1-powered-think-deeper-for-free-232723768.html.

Westerstrand, S., R. Westerstrand, and J. Koskinen. 2024. "Talking Existential Risk into Being: A Habermasian Critical Discourse Perspective to AI Hype." AI and Ethics 4 (3): 713-26.

Winner, L. 1980. "Do Artifacts Have Politics?" Daedalus 109 (1): 121-36.

Zuboff, S. 2020. "You Are Now Remotely Controlled," New York Times, Jan 24.

Research on this essay was funded in part by an award from the Cosmos Institute, and by the Templeton World Charity Foundation.

© 2025, Seth Lazar

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

ENDNOTES

1 This is obviously not a new observation. For earlier discussions of anticipatory ethics, and in particular the challenge of ethical foresight in the face of potentially transformative technologies, see, as well as Johnson (2011), Bostrom (2007); Lucivero et al. (2011); Brey (2012). Drawing a clear line between anticipatory ethics and the ethics of technology more generally is hard, but David Collingridge (1980) seems to have been the first to recognise the particular challenges involved in ethical investigation into technologies that are, in real time, rapidly changing society. A complementary approach has been developed in the social sciences; see the concept note on "Anticipatory Social Research" from Alondra Nelson and David Banks, here: https://www.ssrc.org/wp-content/uploads/2021/12/5fae9aa1f3fe7.pdf.

2 Speculative ethics was originally associated with a rich form of speculation associated with science fiction—for an early critique, see Nordmann (2007). For contemporary discussion of "sociotechnical speculative ethics" that is basically equivalent to anticipatory ethics as described here, see Gabriel et al. (2024: 6).

3 Compare the very similar messages of Nordmann (2007) in response to "speculative NanoEthics" and https://www.nature.com/articles/d41586-023-02094-7 (headline: "Stop talking about tomorrow's AI doomsday when AI poses risks today").

4 Related points are made in many recent articles and books. Here is a sample: Helfrich (2024); Sloane et al. (2024); Vallor (2024); Varoquaux et al. (2024); Westerstrand et al. (2024). For an overview and direct response to some of these arguments, see Swoboda et al. (2025).

5 Nordmann (2007).

6 See also Burgess (2023) on "Big Critique."

7 See Hendrycks et al. (2023).

8 See fn 4 above for references.

9 See Sloane et al. (2024) for a complementary view.

10 Of course, for an individual who has been wrongly arrested, for example, the consequences are severe. However, clearly it is worse for a million people to suffer the same wrong than it is for eight to do so. It is also worth distinguishing harms that are due to discriminatory algorithms from those that are due primarily to background social structures that are discriminatory, where AI makes little or no marginal difference.

11 For example, think of some canonical examples of algorithmic discrimination: Google's racist photo classification, the early search results for the term "black girls," and even differentially accurate facial recognition systems. In each of these cases, discriminatory algorithms were identified early on and were amended or improved as a result. One can debate the magnitude of the harm that ultimately resulted, but in my view, these are all examples of anticipatory ethics working quite well. See e.g., Buolamwini and Gebru (2018); Noble (2018). Thanks to Dean W. Ball for prompting my thinking on this.

12 On the last of these, see Neal Stephenson's The Diamond Age.

13 For example, as Helen Toner reports, even noted AI sceptic Gary Marcus puts human-level AI within two decades; Yann Le Cun within one. See https://helentoner.substack.com/p/long-timelines-to-advanced-ai-have.

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

14 This is true for computational systems more generally, as Johnson (2011) observed.
15 For methods in anticipatory ethics that make use of foresight studies, see Brey (2012); Floridi and Strait (2020).
16 Chan et al. (2023); Gabriel et al. (2024); Kolt (2024); Lazar (2024); Kapoor et al. (2025).
17 Kapoor et al. (2024).
18 Though, of course, more speculative thinking can have its rewards (and in any case, people should write about what they want).
19 Obviously answers to each of these questions admit of degrees.

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

About the Author

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.

Lazar is the Knight First Amendment Institute's 2024-2025 senior AI advisor.

Electronic copy available at: https://knightcolumbia.org/content/anticipatory-ai-ethics

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