Anticipation Is More Than Prediction: Proactively Preventing Harms in AI Development

Jared Katzman
apricity@umich.edu
University of Michigan, Ann Arbor
Ann Arbor, Michigan, USA

Tawanna Dillahunt
tdillahu@umich.edu
University of Michigan, Ann Arbor
Ann Arbor, Michigan, USA

Ben Green
bzgreen@umich.edu
University of Michigan, Ann Arbor
Ann Arbor, Michigan, USA

ABSTRACT

Regulations, corporate governance frameworks, and academic conferences now require technologists to anticipate the societal impacts of the artificial intelligence (AI) tools they develop. These efforts aim to limit the unintended harms of AI systems. However, computer scientists lack the training and methods to accomplish this goal. To help address this challenge, we provide theoretical and practical lessons for how technologists should anticipate the effects of AI throughout the research and development process. By synthesizing insights from the fields of future studies, safety engineering, and anticipatory governance, we argue that AI practitioners need to reconceive what anticipation involves. In contrast to the approach that most computer scientists take, anticipation is not simply the act of predicting the impacts of a new technology. Instead, anticipation involves combining foresight and action to build capacity for mitigating the potential harms of technology. We contribute P-FAGE, a framework that outlines the four essential steps of an anticipatory process: planning, foresight, action & governance, and evaluation. We then highlight five key decisions that AI practitioners should make during the planning step, with recommendations for best practices. We conclude by discussing limitations and suggesting how institutional processes and regulations can encourage AI practitioners to adopt a more rigorous approach to anticipation.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → **Governmental regulations**; **Codes of ethics**; **Socio-technical systems**.

KEYWORDS

anticipation, foresight, AI policy, algorithmic impact assessment, AI safety, anticipatory governance

ACM Reference Format:
Jared Katzman, Tawanna Dillahunt, and Ben Green. 2025. Anticipation Is More Than Prediction: Proactively Preventing Harms in AI Development. In *Proceedings of Workshop on Sociotechnical AI Governance: Opportunities and Challenges for HCI (CHI '25).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

There are many documented instances of artificial intelligence (AI) technologies leading to societal harm that developers neither intended nor expected [4, 22]. In response, scholars [9], policymakers [18, 23, 24], and civil society organizations [5] have called for technologists to proactively assess and mitigate the potential negative impacts of AI before deployment. However, computer scientists struggle to proactively mitigate the harms of AI systems. First, computer scientists often perceive anticipation as an impossible task [3, 14]. They believe that they lack agency to affect downstream impacts, such that negative outcomes will happen regardless of their design choices [12]. Second, even when computer scientists do attempt to anticipate, their efforts tend to be limited in scope. Computer scientists might mention the potential harms of their work, but rarely implement or discuss possible mitigations [12]. Technical researchers focus on a narrow range of probable harms resulting from technical system errors, overlooking possible social impacts technology can have on a broad set of stakeholders [3].

We argue that AI practitioners need to shift their conceptual and methodological approach to anticipation. Computer scientists tend to operationalize anticipation primarily as predicting the possible impacts of a technical system. However, the goal of anticipation is not to predict what will happen in the future with some degree of certainty. Instead, based on its Latin roots—*ante-* (before) and *-capere* (capacity)—anticipation involves combining foresight and action to build a capacity to identify and mitigate the potential harms of technology [8]. Thus, AI practitioners should attempt to foresee a wide range of possible consequences and then proactively take steps to prevent harm from materializing.

To operationalize a more rigorous approach to anticipation into practice, we propose structuring anticipation around a four-step process that we call P-FAGE: planning, foresight, action & governance, and evaluation. In addition, we include a discussion of some initial steps the AI community should take to shape institutional processes and regulations that would facilitate these improved anticipatory practices. Throughout the paper, when we use the term "anticipate," we refer to the combination of these four stages—it is not simply a synonym for "foresee." This framework is part of a larger paper under review where we provide lessons from three fields that have previously theorized how to anticipate the impacts of technological innovations.

2 MORE THAN PREDICTION: ANTICIPATION AS A MULTI-STAGE PROCESS (P-FAGE)

AI practitioners need more actionable guidance for operationalizing a more rigorous approach to anticipation. Thus, we conducted a meta-narrative review by mapping the debates and developments of three fields: future studies [2], safety engineering [7], and anticipatory governance [1]. By closely reading heavily cited articles and recent literature, we identified themes related to (a) how each field operationalizes anticipation into practice and (b) research that critiques anticipatory practices or demonstrates how to improve them. Synthesizing previous frameworks [10, 25], we propose that AI

CHI '25, April 26-May 1, 2025, Yokohama, Japan
2025. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

practitioners can view anticipation as a process structured around four key stages: planning, foresight, action & governance, and evaluation. We call this process P-FAGE. In this acronym, we separate the planning stage from the other three stages to highlight its importance in setting the terms of any anticipatory process. This stage is often overlooked; however, if practitioners engage thoughtfully in the planning stage, they can significantly increase the capacity of their anticipatory processes to promote the benefits and avoid the harms of AI. In our full paper, we outline best practices to improve developers' capacity during the planning stage. We provide a summary of key themes in Table 1.

1) Planning: AI practitioners must first determine the parameters of an anticipatory process. This includes deciding on who to involve in the process, what technical system to analyze, what kinds of impacts to foresee, what methods to use to foresee, and what process they will use to respond [25]. Practitioners should formalize the planning stage, even if not required by external requirements, because they can use it to increase their anticipatory capacity (i.e., the time, resources, and expertise they have to shape the future impacts of technology). For instance, if a group of researchers realizes before writing their broader impact statements that they don't have enough expertise to identify the social implications of technology, they should bring in additional stakeholders to help them foresee more types of impacts and create coalitions with others outside their team to collaboratively address them.

2) Foresight: After planning, practitioners can turn to the activity most commonly associated with anticipation: foreseeing the potential future impacts of a technology. There are a wide range of methods that practitioners can employ to generate statements and scenarios about the future [1, 11, 16]. To inform their choice of methods, practitioners should first determine which future impacts they wish to analyze (e.g., physical vs. social harms). Practitioners also need to decide how they will evaluate the likelihood and desirability of any scenarios they foresee. Technical experts tend to overlook how these choices rely on normative and often political choices of how one defines the future. Therefore, AI practitioners should integrate diverse perspectives, from non-technical experts and affected stakeholders, to expand their foresight capabilities and increase their anticipatory capacity [1].

3) Action & Governance: After conducting foresight exercises, practitioners must work to mitigate potential harms and foster potential benefits. Without this step, "anticipation" is merely a foresight exercise. After identifying scenarios where a system leads to negative impacts, practitioners should take proactive steps to minimize the likelihood and severity of harm. Some negative impacts of technology may be outside the direct control of a development team (e.g., due to organizational structures). In those cases, practitioners should engage other stakeholders to address possible harmful scenarios. If risk cannot be mitigated to a satisfactory degree, then practitioners should consider stopping their development of a system or not deploying an already-developed system. In addition, because it is impossible to eliminate the potential for harm [7], practitioners need to implement governance procedures for responding when negative impacts occur. These governance practices could include conducting regular system evaluations, implementing monitoring systems to catch shifts in model behavior, and developing organizational processes to solicit user feedback about the effects of the system.

4) Evaluation: As a final step, practitioners should evaluate their anticipatory process, seeking ways to improve their anticipation in the future. Because anticipation requires taking actions that influence the future, practitioners should not evaluate whether an anticipatory process was successful based on whether they accurately predicted future outcomes [13]. Instead, evaluations should assess how practitioners conducted foresight exercises, documented their work, and acted on knowledge generated from the process (e.g., did they take appropriate actions based on their expectations of potential harm?) [17]. Evaluations should also consider what strategies practitioners put in place for when harm occurs. Consider a scenario in which practitioners foresaw a potential harm and took action to prevent it. In this case, the forecast would be incorrect in a narrow predictive sense, since the outcome did not occur. However, this would be a successful anticipatory process, since the practitioners used foresight to take actions that led to the desired outcome.

3 INSTITUTIONALIZING A NEW APPROACH TO ANTICIPATION

Adopting new approaches to anticipation is difficult for practitioners, as they must respond to the structures and incentives of their professional communities, employers, and broader innovation economy.

3.1 Implications for AI Institutions

While this paper presents a general framework for anticipating the harms of AI systems, its practical application will necessarily vary depending on the institutional and economic incentives shaping AI development. Our framework is primarily motivated by the challenges that even well-intentioned practitioners face in anticipating and addressing harm. As critical scholarship has shown, good intentions alone are insufficient to prevent AI systems from producing adverse effects on individuals and society [6, 20]. Anticipation, in this context, offers a pathway for integrating broader perspectives into the development lifecycle, enabling greater reflection and responsiveness to epistemic and procedural blind spots [21] The P-FAGE framework advocates for an iterative approach to anticipation, in contrast with prevailing practices that treat it as a one-time checklist. We see this intuition already reflected by some developers. For instance, when researchers at Google implemented Farsight, a tool to help developers foresee possible risks of generative AI models, their users wanted the tool to provide insights for mitigative actions too [26]. Our proposed five key decisions helps a practitioner seeking to increase their critical reflexivity by examining the limitations of their current anticipatory practices and identify opportunities to expand their scope of foresight and responsiveness. For example, when designing a new AI system, practitioners might ask: Are there additional stakeholders who should be included in this process, and how can their perspectives be meaningfully incorporated? If the team is focused narrowly on algorithmic fairness, who might they engage to surface environmental, infrastructural, or long-term societal impacts? Instead of expecting engineers alone to come up with mitigative actions,

Table 1: How Planning Decisions Shape Each Stage of the P-FAGE Process.
VA Case: Suppose that the US Department of Veterans Affairs (VA) is proposing a pilot program where they provide veterans seeking mental health support with a large-lagnauge model (LLM) chatbot mobile app while they wait for a backlogged appointment with a counselor.

Planning Decision	Foresight	Action & Governance	Evaluation
1. Decide Who is in the Room (And Who Is Not)	**General:** Involving marginalized groups and interdisciplinary experts expands the scope of plausible futures and values considered.	**General:** Enables governance actions that reflect pluralistic values, not just technical requirements.	**General:** Measures effectiveness based on whose perspectives were integrated and whether harms to marginalized groups were mitigated.
	VA Case: Including transgender veterans and social scientists surfaces risks like identity-based exclusion from the chatbot and inadequate responses to sensitive topics.	**VA Case:** Co-designed chatbot interventions account for intersectional needs (e.g., LGBTQ+ mental health).	**VA Case:** Evaluation includes whether marginalized users trust actions and influenced iterative updates.
2. Define the Sociotechnical System	**General:** Expanding boundaries to include upstream and downstream effects reveals risks otherwise excluded.	**General:** Supports governance across multiple levels of institutions and systems.	**General:** Evaluates short-term and long-term effects across interconnected systems.
	VA Case: Foresees veteran overdependence on chatbot and reduced participation in human counseling programs.	**VA Case:** Synchronizes chatbot design with VA staff training and policy on app usage.	**VA Case:** Tracks whether app usage patterns complemented or disrupted other VA care services.
3. Specify What Outcomes to Anticipate	**General:** Broadens anticipation beyond technical risks to include systemic, representational, and distributional harms.	**General:** Promotes governance that anticipates secondary impacts and mitigates structural inequalities.	**General:** Specifying the definition of an outcome supports the selection of evaluation metrics.
	VA Case: Surfaces risks like VA justifying cutting in-person services based on high chatbot usage.	**VA Case:** Prevents overreliance on the chatbot as a cost-saving measure, commits to stakeholder-informed policy guardrails.	**VA Case:** Tracks whether the app altered access to care, especially for underserved groups.
4. Select Methods for Foresight	**General:** Mixed methods (e.g., quantitative, participatory, speculative) yield more comprehensive scenarios.	**General:** Enables nuanced mitigation strategies that can respond to multiple possible futures.	**General:** Evaluates which methods generated actionable insights and which were performative.
	VA Case: Combines red-teaming with user workshops to uncover edge-case harms and trust dynamics.	**VA Case:** Creates governance processes that are resilient to changes in the political environment (e.g., protecting data if abortions become illegal).	**VA Case:** Assesses which foresight techniques identified real harms and shaped meaningful improvements.
5. Outline Process for Responding	**General:** Early planning of responses ensures capacity to act, rather than react.	**General:** Creates recourse channels for stakeholders to identify unforeseen harms.	**General:** Evaluates processes according to responsiveness and preparedness.
	VA Case: Anticipates trade-offs of different escalation strategies (e.g., police contact vs. veteran-defined support).	**VA Case:** Creates escalation protocols, VA hospital notifications, and external advisory oversight.	**VA Case:** Measures how effectively harm was addressed through how robust governance protocols are.

researchers should investigate tools for facilitating iterative discussions between diver stakeholders such as engineers, sociotechnical researchers, compliance officers, and impacted communities.

At the same time, we recognize that institutionalizing a forward-looking, care-centered approach to AI development faces significant structural barriers. Chief among these is the political economy of AI, in which prevailing capitalist logics prioritize profitability and efficiency over ethical responsiveness. Rigorous anticipatory processes often require substantial investments of time, labor, and

financial resources—investments that may conflict with the incentives of firms operating under shareholder primacy or competitive market pressures. The externalization of social and environmental costs—such as harms to labor, inequality, or ecological degradation—is a well-documented feature of capitalist production models. As such, even mission-aligned organizations may struggle to meaningfully invest in mitigation efforts when constrained by limited resources and institutional competition. Policy reforms could begin to shift these dynamics by offering material incentives for harm reduction or by increasing liability for failures to conduct adequate anticipatory assessments. Our framework offers a starting point for discussions on such reforms by outlining the components of a rigorous anticipatory process and helping define what might count as "sufficient" foresight.

3.2 Implications for Policymakers

Regulation can play a crucial role in prompting companies and governments to invest in robust anticipatory processes for AI development. One important task for policymakers is to enforce standards for what counts as sufficiently rigorous anticipation. Notably, while the AI Bill of Rights and other policies call for anticipation, they do not clarify the standard for which harms are "reasonably foreseeable" [15, 24]. Our work suggests that policymakers should make this judgment based on a development team's anticipatory process. Policymakers could bolster existing impact assessment requirements by asking developers to describe what methods they used to foresee potential impacts and what actions they took to mitigate risks. If an AI tool leads to harm, policymakers would then scrutinize the development team's practices. Did the team engage in each step of the P-FAGE process? Did they follow best practices for how to conduct those steps? If the answer to either question is no, then policymakers should hold developers liable for failing to address reasonably foreseeable harms. Of course, an important open question is how to develop standards that would help policymakers make these judgments about a development team's anticipatory process.

However, we note that regulation alone is unlikely to suffice, particularly in cases where organizations are openly hostile to social accountability, or where formal compliance mechanisms are co-opted for ethics-washing [19]. While some companies may approach regulation superficially, the resulting documentation trail can serve as a basis for audit and regulatory review, offering traceable records that can be used to evaluate foreseeability claims after harms occur. Yet accountability cannot rest solely with regulators. Thus, while meaningful anticipation will require government oversight, it will also require the redistribution of power and legitimacy across the AI development ecosystem. We argue for the inclusion of civil society actors such as unions, community organizations, and AI governance advocacy groups. These actors should be embedded into liability regimes with legal and institutional mechanisms that empower them to contest industry claims and assert alternative standards of foresight and responsibility. Furthermore, AI practitioners themselves must be granted labor protections, including the right to collective action and protected disclosures, so that they can surface risks without fear of retaliation. Without these structural supports, anticipatory processes will remain vulnerable to institutional inertia and political capture.

REFERENCES

[1] Daniel Barben, Erik Fisher, Cynthia Selin, and David H Guston. 2008. Anticipatory Governance of Nanotechnology: Foresight, Engagement, and Integration. *The handbook of science and technology studies* 979 (2008).

[2] Wendell Bell. 2003. *Foundations of Futures Studies: Human Science for a New Era.* Vol. 1. Transaction Publishers, New Brunswick, N.J.

[3] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416 [cs]

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* PMLR, 77–91.

[5] Eliza Campbell and Michael Kleinman. 2023. AI Must Not Become a Driver of Human Rights Abuses.

[6] Sasha Costanza-Chock. 2020. *Design Justice: Community-led Practices to Build the Worlds We Need.* The MIT Press.

[7] Clifton A. Ericson. 2011. *Concise Encyclopedia of System Safety: Definition of Terms and Concepts.* Wiley, Hoboken, NJ.

[8] David H. Guston. 2013. "Daddy, Can I Have a Puddle Gator?": Creativity, Anticipation, and Responsible Innovation. In *Responsible Innovation*, Richard Owen, John Bessant, and Maggy Heintz (Eds.). John Wiley & Sons, Ltd, Chichester, UK, 109–118. https://doi.org/10.1002/9781118551424.ch6

[9] Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi DeRussis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. 2021. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. arXiv:2112.09544 [cs]

[10] Raija Koivisto, Nina Wessberg, Annele Eerola, Toni Ahlqvist, Sirkku Kivisaari, Jouko Myllyoja, and Minna Halonen. 2009. Integrating Future-Oriented Technology Analysis and Risk Assessment Methodologies. *Technological Forecasting and Social Change* 76, 9 (Nov. 2009), 1163–1176. https://doi.org/10.1016/j.techfore.2009.07.012

[11] Nancy Leveson. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety.* MIT Press, Cambridge, Mass.

[12] David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R. Hullman, and Tina Eliassi-Rad. 2022. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22).* Association for Computing Machinery, New York, NY, USA, 424–435. https://doi.org/10.1145/3514094.3534155

[13] Mihai Nadin (Ed.). 2016. *Anticipation Across Disciplines.* Cognitive Systems Monographs, Vol. 29. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-22599-9

[14] Nassim Parvin and Anne Pollock. 2020. Unintended by Design: On the Political Uses of "Unintended Consequences". *Engaging Science, Technology, and Society* 6 (Aug. 2020), 320–327. https://doi.org/10.17351/ests2020.497

[15] Jason Pielemeier, Ramsha Jahangir, and Hilary Ross. 2024. Ensuring Digital Services Act Audits Deliver on Their Promise. *Tech Policy Press* (Feb. 2024).

[16] Rafael Popper. 2008. How Are Foresight Methods Selected? *Foresight* 10, 6 (Jan. 2008), 62–89. https://doi.org/10.1108/14636680810918586

[17] René Rohrbeck. 2010. *Corporate Foresight: Towards a Maturity Model for the Future Orientation of a Firm.* Springer Science & Business Media.

[18] Manny Rutinel, Brianna Titone, and Robert Rodriguez. 2024. Consumer Protections for Artificial Intelligence.

[19] Andrew D Selbst. 2021. An Institutional View of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology* 35, 1 (2021).

[20] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, Atlanta GA USA, 59–68. https://doi.org/10.1145/3287560.3287598

[21] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a Framework for Responsible Innovation. *Research Policy* 42, 9 (Nov. 2013), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

[22] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. https://doi.org/10.2139/ssrn.2208240

[23] Elham Tabassi. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* Technical Report NIST AI 100-1. National Institute of Standards and Technology (U.S.), Gaithersburg, MD. NIST AI 100–1 pages. https://doi.org/10.6028/NIST.AI.100-1

[24] The White House. 2022. *Blueprint for an AI Bill of Rights.* Technical Report. Office of Science and Technology Policy.

[25] Sergio Urueña. 2023. Enacting Anticipatory Heuristics: A Tentative Methodological Proposal for Steering Responsible Innovation. *Journal of Responsible Innovation* 10, 1 (Jan. 2023), 2160552. https://doi.org/10.1080/23299460.2022.2160552

[26] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. https://doi.org/10.1145/3613904.3642335 arXiv:2402.15350 [cs]