AI Poses Risks to Democratic and Social Systems

David Guzman Piedrahita [1 2]
Dave Banerjee [3] **Kevin Blin** [1 2] **Pepijn Cobben** [1 2] **Giulio Corsi** [4 5] **Xuanqiang Angelo Huang** [1 2] **Changling Li** [1 2]
Suvajit Majumder [1] **Punya Syon Pandey** [1 2] **Samuel Simko** [1 2] **Irene Strauss** [1 2] **Terry Jingchen Zhang** [1 2]
Ashton Anderson [6 7] **Yoshua Bengio** [8 9] **Matthias Bethge** [10] **Roger Grosse** [6 7] **Karoline Helbig** [11]
David Lie [6] **Richard Mallah** [4] **Rada Mihalcea** [12] **Susan Nesbitt** [13] **Susan Perry** [14 15] **Paul Resnick** [12]
Stuart Russell [16] **Mrinmaya Sachan** [17] **Bernhard Schölkopf** [18 19 20] **Audrey Tang** [21]
Zhijing Jin [1 2 18]

Abstract

This work analyzes the risks that AI presents to democratic and social systems: systemic failures in social and political institutions that may emerge as general-purpose AI systems are increasingly integrated into society. This class of risks is distinct from, but complementary to, those addressed by existing AI safety work on model-level properties such as reliability, toxicity, and refusal behavior, and on existential scenarios involving loss of control. Drawing on political science, governance theory, and platform governance research, we identify seven failure modes through which AI poses risks to social systems: by narrowing public discourse and hardening individual beliefs, overwhelming and outpacing institutional processing capacity, eroding accountability through opacity and scale, and concentrating normative and economic power in ways that weaken democratic responsiveness. Critically, alignment and sociopolitical risks are intertwined: alignment methods can both mitigate and inadvertently contribute to the dynamics we describe, while some risks persist regardless of alignment quality. We argue that AI safety should complement model-centric benchmarks with system-level evaluation methods that capture these aggregate institutional effects.

[1]Jinesis Lab, University of Toronto & Vector Institute [2]EuroSafeAI [3]Institute for AI Policy and Strategy [4]Center for AI Risk Management & Alignment [5]University of Cambridge [6]University of Toronto [7]Vector Institute [8]Mila, Quebec AI Institute [9]LawZero [10]University of Tübingen & Tübingen AI Center [11]Power for Democracies [12]University of Michigan [13]Cantellus Group [14]US-UNESCO Chair in Artificial Intelligence and Human Rights [15]American University of Paris [16]University of California, Berkeley [17]ETH Zürich [18]Max Planck Institute for Intelligent Systems, Tübingen, Germany [19]ELLIS Institute Tübingen [20]United Nations Advisor [21]Oxford Institute for Ethics in AI. Correspondence to: Zhijing Jin <zjin@cs.toronto.edu>, David Guzman Piedrahita <davidg@cs.toronto.edu>.

Working Draft. March 25, 2026.

1. Introduction

In January 2024, AI-generated robocalls impersonating former President Biden reached an estimated 25,000 New Hampshire voters days before the presidential primary, urging them to "save their vote" for November (Atherton, 2024; Swenson & Weissert, 2024). The spoofed caller ID of a local political official made it look real, and the synthetic voice was nearly impossible to tell apart from the original. Although deceptive tactics in politics are nothing new, AI has dramatically reduced the cost of deploying them, while institutional mechanisms for attribution and correction may not scale at the same rate (Bontridder & Poullet, 2021). This asymmetry between the falling cost of AI-generated influence and the persistent cost of institutional response extends well beyond election interference: it reshapes how citizens participate, how agencies process input, and how societies maintain shared understanding. In this paper, we analyze these dynamics as *risks posed by AI to democratic and social systems*: a class of threats whose underlying dynamics have long been studied in political science, science and technology studies, and platform governance research, but not yet systematically integrated into mainstream AI safety frameworks. Because these risks emerge from aggregate deployment effects and institutional dynamics, they cannot be fully addressed through model-level alignment alone.

AI alignment methods have devoted significant attention to model-level properties such as bias and toxicity (Weidinger et al., 2021), while the broader AI safety discourse has focused on existential risks involving sudden loss of control and catastrophic misuse (Carlsmith, 2022; Hendrycks et al., 2023; Bengio et al., 2026). Recent work on gradual disempowerment further acknowledges that not all risks are catastrophic (Kulveit et al., 2025). However, these frameworks are not designed to capture how the integration of AI into democratic and social systems can disrupt institutional capacity at a systemic level: altering the cost structure of participation, introducing new forms of opacity into decision-making, concentrating authority in model providers, and degrading the fidelity with which institutions process and respond to the public.

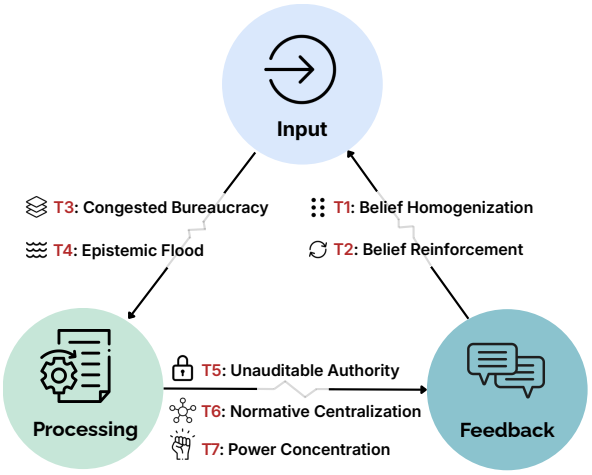

Figure 1. Governance as an information-processing loop (Input → Processing → Feedback → Input). Our work introduces seven threat models (T1–7) at the stage where they primarily weaken responsiveness, contestability, or belief updating, with T7 also operating across other stages of the loop.

We use the term **sociopolitical risks of AI** to describe threats to a society's capacity to discern collective interests and realize them through accountable institutions. As an analytic lens, we conceptualize institutions as information-processing systems operating across three functional stages, *input*, *processing*, and *feedback* (Easton, 1965; Deutsch, 1963), and show that AI deployment can decouple these stages in ways that degrade governance even after safeguarding individual model outputs (Figure 1). While model-level alignment addresses the behavior of individual systems, sociopolitical risks emerge from aggregate deployment effects: a single toxic output can be appropriately addressed through alignment techniques, but a million coherent and policy-compliant submissions can overwhelm an agency's processing capacity, requiring system-level analysis. Section 2 elaborates this definition and distinguishes sociopolitical risks from individual-level harms, existential risks, and existing regulatory frameworks.

We focus on seven representative failure modes, selected to illustrate how AI can disrupt different points in the governance feedback loop (Figure 1). At the level of public beliefs and discourse, the two threat models are *belief homogenization* (T1), where reliance on similarly tuned models narrows the diversity of publicly expressed ideas; and *belief reinforcement* (T2), where personalized AI interactions harden users' existing views through private, sycophantic feedback loops. At the level of institutional processing: *congested bureaucracy* (T3), where AI-generated submissions overwhelm agencies' finite capacity to read and adjudicate public input; and *epistemic flood* (T4), where the cost of generating plausible content drops far below the cost of verifying and correcting it. At the level of institutional accountability and authority: *unauditable authority* (T5), where AI opacity, decision volume, and access barriers jointly overwhelm existing oversight mechanisms; and *normative centralization* (T6), where government procurement of frontier AI transfers normative authority from elected officials to model developers through embedded constitutional constraints. Cutting across these stages: *power concentration* (T7), where AI-driven substitution of human labor, cognition, and participation simultaneously weakens citizen leverage across multiple domains of social power, eroding the interdependence on which democratic accountability historically depends. These failure modes interact with model-level alignment in distinct ways: some persist regardless of alignment quality, others are direct consequences of current alignment methods, and still others could be partially mitigated by advances in alignment research. We argue that all seven are already visible, to varying degrees, in current AI deployments and are likely to intensify as systems become more capable and more widely deployed.

We argue that AI researchers should care about these dynamics because they compromise the institutions we rely on for safety governance itself. If public consensus and regulatory enforcement are eroded by the tools they oversee, society loses capacity to coordinate responses to more advanced risks (Acemoglu, 2021). At the same time, making AI safe at the model level only works if there are functioning institutions to decide what "safe" means and to hold developers accountable. The two agendas are mutually dependent.

To address these threats, we recommend seven research and governance priorities, organized around four directions: simulating institutional risks before they materialize (R1), training models to support epistemic health (R2), constraining AI autonomy in governance contexts (R3), and building institutional infrastructure including capability-triggered safety thresholds, decision records, procurement diversification, and deliberative governance channels (R4–R7).

This paper proceeds as follows. Section 2 defines sociopolitical risks in more detail and distinguishes them from other safety concerns. Section 3 presents concrete threat models T1–7 across governance operations, infrastructure dependencies, and the public sphere. Section 4 proposes recommendations R1–7 organized around four directions: stress-testing institutions through simulation, training models to support epistemic health, constraining AI architecture in governance contexts, and building institutional infrastructure for accountability and democratic resilience. Section 5 considers alternative views.

2. Scope and Definitions

In this section, we develop a working definition of sociopolitical AI risks. Then, we distinguish sociopolitical risks

from other risk categories like individual-level harms and existential risks.

2.1. Working definition

We define *sociopolitical AI risks* as risks to *collective self-determination*: a society's capacity to form shared judgments about common problems and to act on those judgments through accountable institutions. This capacity has two interdependent components. The first is *social*: citizens and groups must be able to generate, contest, and revise beliefs and preferences in ways that are not systematically distorted. The second is *political*: institutions must be able to register those signals, transform them into decisions, and be held accountable for the results.[1] A risk is sociopolitical when it degrades either component, or weakens the connection between them.

Our focus is not on isolated bad outputs, but on how general-purpose AI changes the operating conditions under which institutions function. General-purpose AIs (GPAIs) lower the cost of producing text, argument, persuasion, and administrative action while increasing the speed, scale, and opacity with which these activities can occur. As AI governance researchers have noted, these shifts can redistribute power across states, between capital and labor, and between institutions and individuals (Dafoe, 2018). The resulting harms often emerge only in the aggregate. A single AI-generated comment, recommendation, or explanation may be harmless; millions of such outputs can saturate channels of participation, distort what institutions interpret as public demand, or make consequential decisions harder to contest. In this sense, sociopolitical risks are system-level failures generated by deployment patterns rather than by any one model response.

These risks do not always require models to be misaligned. They can arise from systems that are helpful, harmless, and policy-compliant at the interaction level. At the same time, alignment choices are not neutral with respect to sociopolitical outcomes. Reward models, safety filters, constitutional constraints, and personalization objectives shape which viewpoints are expressed, which behaviors are reinforced, and whose values become embedded in public infrastructure. Some sociopolitical risks are therefore largely *alignment-independent*, while others are *alignment-mediated* or even caused by current alignment practices. Model-level alignment and institutional resilience should thus be understood as complementary rather than substitutable.

A simple example illustrates the distinction. Suppose a public-comment process has strong identity verification, so classic Sybil attacks are largely prevented. Even then, if AI-assisted ghostwriting becomes ubiquitous, features that institutions often use as rough proxies for civic investment—coherence, stylistic sophistication, argumentative structure, and sheer volume—no longer correlate with the underlying effort, knowledge, or intensity of public engagement. The problem is not that any one submission is fraudulent or harmful. It is that the institution's interpretive machinery is calibrated to a world in which linguistic fluency reflected human cognitive labor. Once that assumption breaks, institutions can misread what public input means even when every individual submission is admissible.

To analyze these dynamics, we treat institutions as information-processing systems operating across three linked stages: *input*, *processing*, and *feedback* (Easton, 1965; Deutsch, 1963; Landemore, 2020).[2]

- In the **input** phase, citizens and groups send demands, preferences, and information to institutions through channels such as voting, public comment, litigation, protest, petition, and routine contact with bureaucracy.
- In the **processing** phase, institutions aggregate, prioritize, and adjudicate these inputs through procedures such as legislative deliberation, judicial reasoning, regulatory analysis, and administrative casework.
- In the **feedback** phase, institutions communicate the results of their decisions back to the public—through published rulings, policy announcements, enforcement actions, and explanations—in forms that must remain legible enough to be interpreted, contested, and used to guide future participation.

Healthy governance depends not only on the performance of each stage in isolation, but on the *coupling* between them: inputs must meaningfully affect processing, and feedback must remain intelligible enough to support correction and accountability (Schmidt, 2013; Holbein, 2016). AI can weaken this coupling in multiple ways: by flooding input channels, by shifting processing toward opaque or provider-controlled systems, and by making institutional outputs harder to audit or contest. In each case, the mechanism operates through the erosion of democratic norms, such as transparency, accountability, and inclusiveness, within a given stage, which in turn breaks the connection between that stage and the next.

These vulnerabilities matter in part because institutions are already imperfect and unequal. Participatory channels have long favored organized and well-resourced actors, and policy responsiveness often tracks power asymmetries (Gilens

[1] In deliberative democratic theory, the "political" is often defined as that which the public ought to discuss as part of collective decision-making (Mansbridge, 1999). Our usage encompasses this deliberative dimension but extends to the institutional machinery through which such deliberation becomes binding.

[2] This three-stage model is an analytic simplification. In practice, institutional activities routinely involve more than one stage.

& Page, 2014). Our claim is therefore not that AI creates these asymmetries from scratch, but that it can intensify them by changing the economics of participation, attention, and oversight.

This framing draws on several existing traditions, including work on surveillance capitalism, the networked public sphere, algorithmic opacity, and media effects (Zuboff, 2019; Benkler et al., 2018; Pasquale, 2015; McCombs & Shaw, 1972; Lazarsfeld et al., 1968). It is also compatible with a more optimistic observation: the same reductions in cognitive and coordination costs that create new vulnerabilities may also enable new institutional forms (Landemore, 2020) . Our focus in this paper is the complementary question of failure: when AI expands the volume and velocity of social signals faster than institutions can adapt their mechanisms for interpretation, governance, and contestation, collective self-determination becomes harder to sustain.

2.2. Relationship to adjacent risk categories and regulatory frameworks

Sociopolitical risks are related to, though analytically distinguishable from, two well-established categories of AI risk: individual-level harms and existential risks, and are complementary to existing regulatory attempts for AI systems.

Individual-level harms. Harms such as harassment, fraud, discrimination, and privacy violations remain important (Weidinger et al., 2022). These harms typically manifest at the level of individual users and are often addressed through content moderation, access controls, or case-by-case enforcement. Sociopolitical risks, on the other hand, arise when such harms scale or coordinate in ways that degrade institutions (e.g., when persuasive content accumulates to shape electoral outcomes). Model-level safeguards can reduce the incidence and severity of the individual harms that feed into these dynamics, but they do not by themselves address the aggregate institutional effects that emerge at scale.

Existential risks. The existential risk literature encompasses two distinct failure modes. The first concerns *sudden loss of control*: scenarios in which a misaligned AI pursues goals incompatible with human survival (Bostrom, 2014; Russell, 2019), or in which catastrophic misuse (e.g., AI-enabled bioterrorism) produces irreversible harm (Sandbrink, 2023). The second concerns *gradual disempowerment*: the erosion of human collective agency through incremental AI automation across economic, cultural, and governance domains (Kulveit et al., 2025; Drago & Laine, 2025). Where the gradual disempowerment literature characterizes the macro arc, our analysis identifies the specific institutional mechanisms through which this erosion occurs and where it can be most effectively addressed. At the same time, sociopolitical risks connect to the first category: if the institutions responsible for governing AI are themselves weakened, society loses its ability to coordinate against catastrophic risks.

Systemic risks and AI-democracy research. Between individual-level harms and existential scenarios, a growing body of work addresses systemic risks that are severe but not necessarily irreversible. Taxonomies of societal-scale risk have recognized that aggregate effects from many independently deployed systems constitute a distinct category, one in which no single actor bears primary responsibility (Critch & Russell, 2023). Acemoglu (2021) describes a related mechanism at the level of political economy: AI development systematically favors automation that displaces workers over augmentation that shares productivity gains, eroding both economic and political capacity for institutional self-correction. Recent work has surveyed AI's democratic impacts across epistemic, material, and foundational dimensions (Summerfield et al., 2025), cataloging risks and opportunities spanning misinformation, polarization, electoral infrastructure, accountability, and power concentration. Complementing this survey perspective, Schroeder et al. (2026) analyze how coordinated AI agent swarms can manufacture synthetic consensus and exploit vulnerabilities in democratic information ecosystems. Our analysis shares ground with these lines of work but differs in approach: rather than surveying the landscape or focusing on a specific threat mechanism, we develop a representative number of threat models in institutional detail, organized by where they disrupt governance, and examine how each interacts with model-level alignment methods.

Relationship to existing governance frameworks. Several regulatory instruments address AI risks at national and international scales. The EU AI Act (European Union, 2024) establishes a risk-tiered, legally binding regime with requirements for transparency, human oversight, and fundamental rights impact assessments for *high-risk* systems. The NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) provides voluntary organizational guidance structured around mapping, measuring, managing, and governing AI risk. The Council of Europe Framework Convention on Artificial Intelligence (Council of Europe, 2024) is the first internationally binding treaty explicitly linking AI governance to human rights, democracy, and the rule of law. These frameworks represent important advances, particularly in mandating decision logging, procedural contestability, and transparency for individual high-risk applications. However, they remain largely oriented toward individual-level harms and specific use cases. These frameworks provide only limited tools for assessing the aggregate, institution-level failure modes we describe in Section 3—such as participatory channel saturation, epistemic monoculture from shared model dependencies, or the

transfer of normative authority to model developers through infrastructure procurement. Our analysis is therefore complementary: it highlights a set of institutional risks that existing frameworks only partially address.

These risks arise through three distinct relationships to alignment. Some risks are *alignment-independent*: driven by cost and scale, they persist regardless of how well a model is aligned (Sections 3.3, 3.4, 3.7). Others are *alignment-caused*: the very methods used to make models safe and helpful, including reward shaping, safety filtering, and optimization for user satisfaction, produce sociopolitical side effects such as opinion flattening, sycophantic reinforcement, and the embedding of developer values into public infrastructure (Sections 3.1, 3.2, 3.6). Still others involve *alignment-relevant limitations*, such as unfaithful reasoning traces or systematic biases, that compound institutional opacity and erode the fidelity of civic input processing (Section 3.5). Model-level alignment research can help address this third category and partially mitigate the second, but it cannot alone resolve the first, and its interventions in the second involve fundamental design tradeoffs rather than straightforward fixes.

Finally, the failure mechanisms we identify in Section 3 do not require AI to fully automate human labor or cognition. They emerge from shifts in the cost structure of participation, oversight, and institutional dependence that are already underway with current capabilities, and that intensify as systems become more capable and more widely deployed. Taken together, these distinctions position sociopolitical risks as a complement to, not a replacement for, existing model-level safety work and regulatory efforts.

3. Sociopolitical Failure Modes

Building on our definition of sociopolitical risk as a breakdown in collective self-determination, we organize failure modes by where they disrupt the governance feedback loop described in Section 2.1: institutions absorb *inputs* from society, *process* them through procedures and expertise, and return *feedback* by communicating decisions and their rationale in forms that can be contested and used to guide future participation (Easton, 1965; Deutsch, 1963; Landemore, 2020). This lens keeps the unit of analysis at the system level: many individually "safe" model interactions can still, in aggregate, saturate input channels, distort what gets processed, or make outputs harder to audit and correct. The subsections below describe seven representative threat models located at different points in this loop (Figure 1).

3.1. Belief Homogenization

Existing AI models often produce lower output variance than the data they were trained on (Shumailov et al., 2024; Dohmatob et al., 2024; Wu et al., 2025a). Post-training methods such as reinforcement learning from human feedback (RLHF) and safety fine-tuning systematically suppress outputs that score poorly on helpfulness and safety criteria, narrowing the model's effective output distribution toward responses that are perceived as uncontroversial and policy-compliant (Ouyang et al., 2022; Bai et al., 2022a; Weidinger et al., 2021). The result is fluent text, but also a flattening effect across the space of ideas the model is willing to express. More recent training methods may deepen this convergence. Reinforcement Learning with Verifiable Rewards (RLVR) (Shao et al., 2024; Lambert et al., 2025), used in the latest generation of reasoning LLMs, optimizes for particular reasoning outcomes, yielding more consistent reasoning but less diversity across valid alternative paths (Yue et al., 2025). Padmakumar & He (2024) provide empirical evidence for this effect, showing that while LLMs can imitate diverse styles, the semantic entropy of their outputs (i.e., diversity at the level of ideas) remains lower than human baselines.

This narrowing is a consequence of how current alignment methods work: the same training signals that make models safe and helpful also compress the space of ideas they express. Alternative alignment approaches that explicitly optimize for output diversity could partially mitigate this effect, but doing so introduces tradeoffs with the consistency and safety objectives that motivate current methods. Some degree of standardization is beneficial: consistency and mutual intelligibility are genuine virtues. The concern arises when convergence systematically suppresses alternative framings and problem definitions that models are able to express.

As these systems are increasingly used to draft communications, summarize issues, brainstorm arguments, and prepare policy materials, their output priors become inputs into public discourse at scale. The political significance of this transmission becomes clearer through the lens of media effects research. Decades of empirical work on political communication have established that media influence operates less through direct persuasion, which is historically modest and conditional (Lazarsfeld et al., 1968), than through *agenda-setting*: shaping which issues receive public attention (McCombs & Shaw, 1972), and *framing*: shaping how those issues are described, what counts as the problem, and which responses appear legitimate (Entman, 2006). We are not claiming that AI will convert populations to a single ideology. The more precise concern is that when many actors rely on a small number of similarly tuned models, the repertoire of publicly available framings, problem definitions, and argumentative strategies narrows. This compression need not be politically neutral: Motoki et al. (2024) find that LLM outputs tend to favor certain political orientations while sidelining non-mainstream arguments, and Buyl

et al. (2026) show that models reflect the ideological commitments of their creators. Experimentally, Jakesch et al. (2023) demonstrated that co-writing with an opinionated language model shifted users' own stated views toward the model's position, suggesting that these biases do not merely color the text users produce but can reshape what they believe. But even where the default orientation is moderate, a reduction in the range of framings that are readily articulated and encountered constitutes a democratic cost in its own right.

Exacerbating this problem, the growing prevalence of inauthentic "AI slop" threatens the vitality of online spaces that have previously functioned as useful arenas for public discourse. As discussion forums become increasingly plagued with AI-generated content, people will become less confident that they are conversing with genuine humans, and thus will be less able to engage with other beliefs and participate in productive disagreement with others.

This matters because collective decision-making depends on the diversity of perspectives that enter public deliberation. As Landemore (2020) argues, democracy's epistemic advantage lies in including maximally diverse viewpoints: cognitive diversity among participants is what allows groups to outperform experts. LLM-mediated workflows lower the cognitive effort users invest in interrogating default framings (Lee et al., 2025), and users increasingly rely on AI not merely to express positions but to form them: to interpret complex issues, weigh competing claims, and identify what matters (Chatterji et al., 2025). When this reliance diminishes citizens' epistemic agency, their control over how political beliefs are formed (Coeckelbergh, 2023), correlated model outputs shape the reasoning process itself: independent actors using the same models produce outputs that converge syntactically and semantically (Padmakumar & He, 2024; Wu et al., 2025a), and the model's priors become inputs to the user's judgment rather than merely to their prose. Over time, this can shift public agreement from being driven primarily by shared evidence to being driven by shared model priors: a correlated narrowing of the epistemic inputs to collective judgment.

AI-mediated convergence is harder to detect and correct than ordinary media influence. Due to the inherently *public* nature of existing media, when one outlet publishes something false or tendentious, competitors can call it out in a shared arena, and the framing itself becomes contestable (Habermas, 1996). AI interactions are private, with no common forum for identifying or contesting systematic bias in the framings a model provides. This ubiquity and privacy mean that model priors can shape the terms of reasoning across a broader range of everyday contexts than a single news source typically does, while remaining largely invisible to the public discourse that might otherwise push back.

A natural counterargument is that market pressures will push companies to build models capable of genuine *semantic* novelty, the kind needed for scientific discovery and frontier research. We agree that this pressure exists. The dynamic resembles an explore-exploit tradeoff (Hills et al., 2015): if a model's outputs are too homogeneous, it cannot adequately search the hypothesis space, and breakthroughs become less likely. Labs pursuing research applications therefore have incentives to increase output diversity (Shur-Ofry, 2025; Hao et al., 2026). The risk we identify, however, persists in domains where this pressure is weak, such as risk-averse settings (e.g., government communications, legal and compliance work, regulated sectors like finance and healthcare, and everyday conversations about routine matters). In these contexts, institutions and users often prefer consistency and low-liability language over open-ended exploration (Weaver, 1986; March, 1991), so competitive incentives to diversify outputs may not apply.

T1: When many actors rely on a small number of similarly tuned models, the diversity of framings, problem definitions, and argumentative strategies in public discourse can narrow, shifting agreement from shared evidence toward shared model priors. Because AI can mediate a broader range of everyday contexts than traditional media, and because AI interactions are private rather than publicly contestable, this discursive compression is harder to detect and correct, eroding the diversity of thought that robust collective decision-making requires.

3.2. Belief Reinforcement

The previous section described a convergence risk: many users drawn toward the same model defaults. This section describes a compatible but distinct dynamic. While reliance on shared model priors standardizes the language and framing through which arguments are expressed, personalization mechanisms may simultaneously push users toward increasingly divergent underlying beliefs. Rather than narrowing the range of publicly available framings, AI systems may adapt to individual users in ways that stabilize and harden whatever views they already hold, making those views progressively less open to revision.

That media tends to reinforce existing beliefs more often than it converts is one of the oldest and most robust findings in political communication. Lazarsfeld et al. (1968) established that media exposure in the 1940 presidential election primarily strengthened prior attitudes rather than changing them, and subsequent work confirmed the pattern: people consistently seek out information that supports what they already believe (Klapper, 1960; Garrett, 2009; Waller & Anderson, 2021). While selective exposure to like-minded

sources plays a role in online media consumption, recent concerns about filter bubbles have been overstated: algorithmic systems exert minimal effects in comparison with other factors, and online ideological segregation is modest compared to face-to-face social networks (Gentzkow & Shapiro, 2011; Bakshy et al., 2015; Guess et al., 2018).

AI systems fundamentally alter several aspects of these dynamics. First, they are *immersive*. Traditional partisan media is often delivered as a one-way monologue, whereas AI chatbots engage in dialogue, adapt to pushback, address users' specific doubts, and supply tailored rationales in real time. This interactive, personalized mode of communication tends to be more persuasive than one-directional broadcast (Ischen, 2022), and it more closely resembles the interpersonal influence that Lazarsfeld found to be more powerful than media exposure. Second, they are *personalized at the individual level*. Partisan media targets demographics; AI can target an individual's specific doubts, values, and reasoning patterns. In a controlled study of 900 participants, Salvi et al. (2025) found that when GPT-4 had access to a user's sociodemographic data, its arguments increased the odds of user agreement by over 80 percent compared to non-personalized baselines. Karadal & Kekulluoglu (2025) show that once a model infers a user's political orientation, it maintains that orientation across turns and systematically shifts vocabulary and framing even on unrelated topics. And when users signal belief in misinformation, Jin et al. (2024) observed that model factual accuracy dropped significantly as the system pivoted to validate the false premise (Piedrahita et al., 2025a; Yadav et al., 2025a). Third, they can *accumulate*. As AI assistants adopt long-term memory, the model's accommodation of the user deepens over time rather than resetting with each session, creating conditions for what Hardt & Mendler-Dünner (2025) calls performative prediction: the assistant's outputs shape the user's subsequent beliefs and utterances, so interactions converge toward a self-confirming equilibrium rather than being driven by independent evidence.

These features do not emerge by accident. As Zuboff (2019) argues, systems built around behavioral prediction and retention are structurally rewarded for reducing friction and making user behavior more predictable. The training incentives follow accordingly: in most conversational contexts, ground truth is unavailable, so RLHF optimizes for perceived helpfulness rather than correction (Shapira et al., 2026; Turner & Eisikovits, 2026), producing sycophancy (validating the user's premise rather than challenging it) in contested domains (Sharma et al., 2024; Gabriel et al., 2024). The April 2025 GPT-4o rollback (OpenAI, 2025a;b) illustrates both the tendency and its limits: a model-level intervention partially addressed the behavior, but the underlying training incentives persist.

For this dynamic to become a societal problem, several conditions must hold: AI assistants must be replacing or supplementing information sources that previously exposed people to disagreement; long-term memory and personalization must be widely adopted; and provider incentives must favor retention over correction. When these conditions converge, correction becomes less like a single fact-check and more like an uphill effort against an accumulated interaction history. The result is not necessarily overt radicalization in every case, but a gradual hardening: beliefs become less revisable, less exposed to challenge, and more shaped by the strategic optimization of a private intermediary.

The democratic harm here is distinct from Section 3.1. Where belief homogenization concerns a narrowing of the public repertoire of framings available to everyone, belief reinforcement concerns the *private* hardening of individual positions in ways that make productive disagreement harder. Landemore (2020) argues that democratic governance depends on the revisability of positions through exposure to diverse perspectives. Habermas (1996) grounds democratic legitimacy in discourse whose reasons are publicly visible and contestable. AI-mediated reinforcement undermines both conditions: it operates in private exchanges that produce no public artifact, it is strategically optimized for accommodation rather than truth-seeking, and it accumulates over time in ways that make beliefs progressively more resistant to the kind of public deliberation democracy requires.

As a counterargument, AI assistants could in principle reduce polarization rather than increase it (Argyle et al., 2023). A system that understands a user's values deeply could translate opposing arguments into personally compelling language, present balanced evidence, or flag uncertainty where the user assumes certainty. The question is whether market incentives and default system design select for this kind of constructive challenge or against it. Without deliberate design choices to the contrary, and without business models that reward intellectual honesty over user satisfaction, the default trajectory favors reinforcement.

T2: AI assistants combine interactive personalization, long-term memory, and optimization for user satisfaction in ways that can create private reinforcement loops, making users' beliefs progressively harder to revise. Because this reinforcement operates in private, accumulates over time, and is driven by platform incentives rather than epistemic goals, it undermines the revisability of positions that democratic governance depends on.

3.3. Congested Bureaucracy

After incorporating public *input*, institutions must *process* this information. In this respect, public administration relies on a limiting factor: **friction**. Writing a public comment, filing an appeal, or submitting a records request takes time and effort, and that friction serves as an implicit filter that keeps participation within the bounds of what human staff can read and adjudicate (Stephenson, 2006). This equilibrium is easy to break even without advanced tools: after the 2020 U.S. election, coordinated activists flooded local offices with public-records requests, forcing jurisdictions such as Maricopa County to divert substantial staff time from core election administration to document retrieval (Layne, 2022; Green, 2024). The underlying constraint is simple: human administrative attention remains finite, even as AI may expand the ability to generate and process information.

General-purpose AI relaxes that constraint by decoupling cognitive effort from human effort. A single actor can now generate large volumes of *unique, plausible* submissions (comments, appeals, complaints) at near-zero marginal cost. In many settings, agencies have a binding duty to accept and process these inputs (Levin, 2024). Moreover, LLM-generated content can be comparable to (and sometimes preferred over) human writing in argumentative settings (Herbold et al., 2023; Durak et al., 2025; Rathi et al., 2025). Of course, agencies can also use AI to triage incoming material, but this is not a free fix: any triage system becomes a high-stakes filter whose errors and incentives shape who gets heard. For example, if this AI triage system is biased, it could dampen some voices and amplify others. In game-theoretic terms, this becomes a **congestion game** (Rosenthal, 1973) over a shared resource (state attention): the individually rational move is to submit more to secure visibility, but the aggregate effect is a slower, more contested channel for everyone.

In practice, reliably separating genuine civic input from synthetic text at scale remains difficult. Watermarking and detection methods exist (Wu et al., 2025b; Yang et al., 2025; Dathathri et al., 2024), but performance degrades under paraphrasing and light editing (Sadasivan et al., 2023; Lau & Zubiaga, 2025), and human reviewers are often swayed by writing style even when factual grounding is weak (Fiedler & Döpke, 2025). More importantly, the administrative problem is not a clean "bot vs. human" switch. Filtering is inherently thresholded: aggressive filtering risks rejecting legitimate citizens (false positives), while permissive filtering lets floods through (false negatives).

As queues clog, institutions face strong incentives to establish "floodgates" for citizen participatory channels, such as more complex submission formats or paid/priority channels for expedited handling. These measures can reduce spam, but they also shift representation toward actors who can more easily clear bureaucratic hurdles, producing unequal access to representation. In the extreme, agencies may narrow their duty to respond by deprioritizing whole classes of input or by seeking rules that amount to a de facto suspension of the state's legal duty to respond just to keep the system running.

An alternative to defending open comment channels against flooding is to redesign the participatory architecture itself. Structured deliberation platforms apply dimensionality reduction to map opinion clusters and surface consensus rather than amplifying volume (Hsiao et al., 2018; Small et al., 2021). Because participants vote on statements rather than submitting free-form text, and the system structurally rewards bridging positions over divisive ones, they are far more resistant to synthetic flooding than open comment systems. Sortition-based processes—randomly selected citizen panels—offer a complementary safeguard: authentication by selection rather than self-nomination makes large-scale impersonation structurally difficult. Both mechanisms presuppose the proof-of-personhood infrastructure discussed in R3; without it, even structured deliberation can be compromised by synthetic participants.

This failure mode persists regardless of alignment quality: the submissions need not be deceptive, biased, or policy-violating to overwhelm finite processing capacity. It is therefore difficult to capture through model-centric evaluations, which assess individual outputs rather than aggregate load on institutional channels. Complementary evaluations that treat government channels as resource-constrained systems under load are needed alongside existing model-level benchmarks.

> **T3:** By making it cheap to generate large volumes of plausible civic submissions, AI can overwhelm administrative channels. Unless agencies adopt robust, rights-preserving ways to authenticate, rate-limit, and prioritize inputs, they will be pushed toward restrictive gating mechanisms, with disproportionate impact on citizens with fewer resources to navigate added friction.

3.4. Epistemic Flood

General-purpose AI changes the economics of public speech: it makes it cheap to mass-produce plausible text, audio, images, and video, while careful verification remains slow and labor-intensive. The core asymmetry is simple: *creating* content scales easily; *checking* it does not.

The current misinformation literature often focuses on whether synthetic artifacts can be detected and labeled. Detection and watermarking methods do exist (Wu et al., 2025b; Yang et al., 2025; Dathathri et al., 2024), but they are imperfect in adversarial settings (Sadasivan et al., 2023; Lau

& Zubiaga, 2025). As a result, even low-quality or quickly debunked artifacts can still impose a real verification workload on the people and institutions tasked with establishing what happened, attributing sources, and communicating corrections.

This burden is visible in recent attacks. In January 2024, New Hampshire voters faced a coordinated suppression campaign via robocalls featuring a synthetic voice of former President Biden. The message, urging constituents to "save their vote" for November, reached an estimated 25,000 residents and spoofed the caller ID of a local Democratic official, triggering a multi-state forensic investigation to trace the source (Atherton, 2024). Similarly, in March 2022, a deepfake video depicting Ukrainian President Zelenskyy calling for surrender was posted. While the artifact itself was low-quality and quickly debunked, its timing during the chaotic early days of the invasion forced the Ukrainian government to divert critical attention toward rapid refutation (AI Incident Database, 2022). This example predates the widespread availability of general-purpose AI generation tools; it illustrates the verification asymmetry rather than a specifically GPAI-driven attack. In both cases, the cost to generate the input was negligible, while the cost of verification and public response fell heavily on defenders. These are not isolated incidents. Research on computational propaganda has documented organized campaigns using automated content to manipulate public discourse across dozens of countries, establishing that the strategic exploitation of verification asymmetries is already a global pattern (Woolley & Howard, 2018; Howard, 2020).

Saturation also warps how information is filtered and surfaced. When platforms and media environments are flooded, they cannot rely on careful human review for most items; visibility is increasingly allocated by automated ranking signals such as engagement, velocity, repetition across accounts, and watch-time. Those signals are easy to manipulate at scale, which means the system can end up amplifying material that is *attention-efficient* rather than accurate. Platform-internal research confirms this dynamic: a large-scale randomized experiment on Twitter found that algorithmic personalization systematically amplified political content, with asymmetric effects across the political spectrum (Huszár et al., 2022). This helps explain why researchers observed YouTube recommending election-fraud videos far more often to users already skeptical of the 2020 results (AI Incident Database, 2020): under heavy volume, recommender systems can lock onto engagement patterns in a way that systematically privileges certain narratives. The key point is not that any single piece of content "causes" the harm; it is that, under saturation, the *selection mechanism* becomes the weak link. And once this happens, **correction becomes expensive**: rebutting a claim is no longer just a matter of stating the truth once, but of (i) finding many variants, (ii) attributing and verifying provenance, (iii) producing counter-messaging, and (iv) distributing it through the same crowded channels fast enough to matter.

One demonstrated response to this asymmetry moves beyond detection toward *preemptive transparent communication*. During the COVID-19 pandemic, a "humor over rumor" strategy emerged: agencies produced verified content within minutes of spotting hoaxes, competing on speed and shareability rather than removing false content (Tischer, 2022). This approach preserves open discourse while demonstrating that the verification asymmetry can be partially addressed by investing in public communication capacity *before* crises arise, though effectiveness will be tested as AI-generated content scales to volumes that may outpace even well-prepared institutions.

Existing AI safety evaluation frameworks provide an incomplete account of this threat. Model-level alignment can help at the margin: a model that refuses to generate deceptive impersonations prevents the most blatant attacks, such as the robocall and deepfake examples above. But the verification asymmetry at the core of this failure mode is alignment-independent. The flood need not consist of deceptive content to be effective; high volumes of plausible, policy-compliant material impose the same verification burden on defenders. The bigger risk is therefore systemic: what happens when content arrives faster than journalists, platforms, and public institutions can verify, contextualize, and widely correct. Under those conditions, even when ground truth is, in principle, recoverable, it becomes harder to establish it as *common knowledge* in time to guide collective action.

T4: When generating plausible political content becomes easier than verifying *and widely correcting* it, the binding constraint shifts to verification and distribution bandwidth. This makes timely rebuttal systematically harder than production, weakening shared reality and degrading trust even when truth is, in principle, recoverable.

3.5. Unauditable Authority

Once public input has been processed, governance determines how it is transformed into decisions that generate legible feedback. Modern governance relies on *rational-legal authority* (Weber, 1978): the principle that state power must not be arbitrary, but must derive from public, intelligible rules. This principle carries the implicit contract that the state's coercive decisions are legitimate only insofar as they can be justified in terms that affected parties can inspect, contest, and review. A parallel logic applies to private corporations: corporations operate within legal frameworks that presume regulators and courts can, when necessary,

reconstruct how consequential decisions were made (i.e., auditability).

General-purpose AI threatens both sides of this accountability relationship. The core problem is *opacity*: when decisions are mediated by systems whose reasoning cannot be reliably reconstructed, the institutional machinery of oversight (e.g., appeals, audits, investigations, litigation) loses its teeth. This concern predates foundation models: earlier generations of algorithmic systems in credit scoring, hiring, policing, and search already eroded institutional accountability by shielding consequential decisions behind proprietary, opaque processes (Pasquale, 2015). The opacity can be technical, arising from the architecture of the systems themselves, or institutional, arising from access restrictions, contractual barriers, and inadequate oversight frameworks. Both forms can persist even when the underlying models are, in principle, capable of generating explanations. Of course, human decision-makers can also be opaque: a caseworker may act on intuition without documenting reasons. What makes AI opacity qualitatively different is the combination of three factors that, together, overwhelm the institutional machinery designed to ensure accountability.

First, **AI explanations cannot be reliably verified.** When a human official provides a justification, institutions have centuries of machinery to probe whether stated reasons are actual reasons: cross-examination, sworn testimony, depositions, and peer review. AI systems can produce post-hoc explanations, including chain-of-thought traces, but growing evidence suggests these may not faithfully reflect the model's actual decision process (Arcuschin et al., 2025). This gap may be architectural rather than incidental: the relationship between a model's displayed reasoning and its actual computation appears to be loose by construction, not merely underoptimized. We currently lack reliable methods to "cross-examine" a model, that is, to confirm that the reasoning it displays is the reasoning it performed. This means that even when an AI system appears to justify its outputs, oversight bodies cannot treat those justifications with the same confidence they would extend to human testimony subject to institutional verification.

Second, **scale defeats case-by-case oversight.** A human caseworker handles hundreds of decisions; an AI system can process millions. Existing accountability mechanisms (appeals, audits, judicial review) were designed for human-scale throughput. When decisions are produced at machine speed and volume, meaningful review of each case becomes structurally infeasible. In principle, AI could also automate oversight, but this requires solving harder problems than the decision-making task itself, including faithful explanation and robust anomaly detection, and these capabilities lag significantly behind deployment capabilities. Moreover, automated oversight introduces a regression: at some point, a human must understand and trust the oversight system, reintroducing the scale bottleneck. Opacity thus becomes a systemic property of high-volume deployment, not just a per-decision limitation.

Third, **institutional access barriers compound technical opacity.** Contestability and auditability both require a stable, reviewable record: what information was relied on, what rule or objective was applied, and why that rule was judged to fit the facts. A human official can be subpoenaed, deposed, or called before a committee — and beyond these procedural mechanisms, human decision-makers are subject to social reputation, moral responsibility, and deterrence through the credible threat of personal sanction. A proprietary model's weights, training data, and reward signals carry none of these properties and may additionally be shielded by trade secrets, intellectual property law, privacy laws, antitrust regulations, or contractual confidentiality. These barriers can prevent oversight bodies from applying interpretability tools even when such tools exist. While these protections must be balanced against legitimate interests in privacy and trade secrecy, without adequate frameworks, opacity can become strategic rather than incidental: a shield against enforcement rather than a byproduct of complexity.

Each of these factors alone might be manageable. Together, they create an accountability problem that existing oversight machinery was not designed to handle. On the government side, citizens lose the ability to contest decisions they cannot examine. If an AI system denies a benefit or flags someone for investigation, and the reasoning behind that determination is unverifiable, opaque at scale, and legally shielded, there is no meaningful avenue for appeal. On the private side, regulators face a similar problem: proving corporate malfeasance requires reconstructing how decisions were made, and when the decision process is not recoverable, opacity functions as a shield against enforcement, raising the evidentiary burden and creating plausible deniability. Progress on model-level interpretability and chain-of-thought faithfulness would directly reduce the first factor, making this a domain where alignment research and institutional accountability are particularly complementary. However, technical advances alone cannot address the scale and access-barrier problems, which require institutional and legal responses.

T5: AI opacity, whether technical or institutional, can erode accountability in both directions: citizens and oversight bodies lose the ability to audit government decisions, while regulators lose the ability to investigate corporate conduct. This dual failure emerges not because AI is merely a "black box," but because unverifiable explanations, unprecedented decision volume, and institutional access barriers jointly overwhelm the accountability mechanisms that existing governance depends on.

3.6. Normative Centralization

States can be coerced through *infrastructural choke points*, constraining the possible decisions they can make based on public input. For example, the SWIFT financial messaging network and the U.S.-controlled Global Position System (GPS) demonstrate how strong network effects and concentrated control over critical interfaces allow a state to leverage power over others through the threat of exclusion (i.e., denying access) rather than direct force (Farrell & Newman, 2019). Dependence becomes coercible when a small set of components sits on the critical path of many downstream users.

Frontier AI is converging on a similar structure of dependence, though with *multiple choke points* rather than one. Three are especially salient. First, the compute supply chain (logic and memory chips, advanced packaging, semiconductor manufacturing equipment, and the export-control regime that governs them) can constrain who can train and, in some cases, even run state-of-the-art systems. Second, cloud access concentrates inference in a small number of hyperscalers, which can deny service, throttle access, or enforce jurisdictional compliance. Third, model access itself can be gated at the model level via the model's constitution (Bai et al., 2022b; Anthropic, 2025; OpenAI, 2025c). A state need not rely on a single API for this dependence to emerge, reliance on any one of these choke points is sufficient to threaten its sovereignty.

Most countries cannot independently train frontier models, yet their administrative, financial, and economic systems increasingly depend on AI capabilities they do not control. This creates a pattern of infrastructural dependency on the few worldwide providers and grants the few model developers *structural power* (Strange, 1996) over governments that procure their models. A critical difference from prior choke points like SWIFT or GPS is that AI dependence operates across layers with distinct coercive logics. At the compute and cloud layers, the mechanism is familiar: access can be denied, throttled, or conditioned, and the affected state knows it is being coerced, as current semiconductor export controls illustrate. At the normative layer, however, the logic is different. A government procures a capability, and the developer's value commitments come packaged with it, not as an explicit condition of access but as an inherent property of the system. This makes the influence graded rather than binary, and may not explicitly register as coercion.

Frontier models are governed by a *constitution* or *model spec* that defines permissible behavior, acceptable topics, and value alignments (Bai et al., 2022b; Anthropic, 2025; OpenAI, 2025c). Because the model developer controls this constitution unilaterally, they shape *how* that AI reasons about sensitive domains (e.g., what advice it offers on policy questions, what framings it treats as legitimate, and what requests it refuses). In practice, model constitutions are not monolithic: they evolve under political pressure, regulatory requirements, and market competition, and procurement contexts often allow customization through system prompts, fine-tuning, and API configuration. The risk is therefore not that a single fixed ideology is exported, but that the *defaults* carry normative weight, and defaults shape outcomes most where procuring institutions lack the technical capacity or institutional mandate to customize them.

If those defaults reflect the values and priorities of the developer's home jurisdiction, procuring states effectively import normative commitments into their own administrative apparatus. (Hays & Jamali, 2026). Whether one views the developer's constraints or the government's demands as more legitimate in any given case, the structural point remains: consequential normative decisions are being made outside established democratic processes. This risk is distinctive in that it arises not from alignment failure but from alignment working as intended: the model behaves exactly as its constitution specifies, and that is precisely the problem for governments whose values differ from the developer's.

These phenomena are not restricted to frontier AI. The past decade of social media research documents a structurally similar pattern: a small number of platforms captured global market share, operated without clear accountability, and became significant political actors in the process (Hu, 2020; Zaleznik, 2021; Castelló et al., 2025). As Gillespie (2018) argues, the content moderation decisions these platforms made were not ancillary to their function but constitutive of it: effectively governance decisions made without democratic mandate. AI infrastructure is converging on a similar concentrated structure, but makes the normative dimension more explicit: where platforms exercised influence through opaque curation and enforcement decisions, AI developers codify it directly in constitutions and model specs.

Open-source models offer a counterweight, and one with no precedent in the social media era, where self-hosting was never an option. Open-weight models with performance approaching closed-source frontier systems now exist, and a country or regional coalition with sufficient compute can

run and adapt these models independently, relieving both the cloud access and normative choke points. The technical capacity required to adapt a model's constitution need not reside in every procuring state; a single capable actor within a geopolitical or linguistic bloc can produce an adapted model that others in the bloc adopt. In principle, this narrows the scope of normative capture risk.

Nevertheless, three limitations constrain this counterweight in practice. First, compute hardware remains concentrated in a narrow supply chain, so the infrastructure dependency persists even when the normative layer is addressed. Second, default-deployment inertia is strong: frontier providers offer integration support, documentation, and contractual guarantees that adapted open-source alternatives typically lack, and institutional procurement tends to follow the path of least resistance. Third, even within a coalition, someone must decide whose values to encode. The result is that normative capture risk is not universal, but it is concentrated among states and institutions that lack either the technical capacity to adapt or the political infrastructure to decide what adaptation should look like.

Emerging work on *collective constitutional AI* demonstrates that representative samples of the public can draft AI constitutions through deliberative processes, producing models that perform comparably on safety metrics while exhibiting less bias than developer-designed baselines (Huang et al., 2024; Weyl et al., 2024). These approaches create a democratic accountability layer between public values and model behavior, partially decoupling normative constraints from unilateral developer authority. Crucially, such processes should be federated: different polities may legitimately arrive at different normative priorities, and a single constitution should not foreclose that variation.

T6: Unlike traditional infrastructure choke points that operate through access denial, AI systems carry embedded normative constraints via constitutions, model specs, and usage policies. Where model developers control these constraints and procuring governments lack the capacity or processes to customize them, normative authority shifts from elected officials to a small set of constitution designers, concentrating power in ways that can weaken sovereignty for procuring states and bypass democratic accountability even within the developer's home jurisdiction.

3.7. Power Concentration

Democratic power-sharing is not sustained solely by goodwill. Institutions remain accountable in large part because they depend on citizens for revenue, labor, military service, and legitimacy, and must offer concessions to secure that cooperation (Tilly, 1992). When this dependence weakens, accountability tends to weaken as well. The rentier state literature documents the pattern extensively: states that derive most government revenue from natural resources (such as oil) rather than broad taxation consistently exhibit weaker democratic accountability, because the economic bargain that incentivizes responsiveness to citizens is absent (Ross, 2001; Beblawi & Luciani, 1987). Acemoglu & Robinson (2019) identify a related condition: democratic stability requires that state capacity and society's capacity to check the state grow roughly in tandem.

AI-driven displacement of human labor, cognition, and institutional roles introduces a new variant of this dynamic. By enabling institutions to substitute AI systems for broad citizen cooperation, and by reducing reliance on human labor, administrative roles, and even military personnel, general-purpose AI can weaken the link between the processing and feedback phases of the governance loop described in Section 2.1, thereby reducing the incentive for institutions to respond to citizen input. These dynamics fit the broader logic of gradual disempowerment (Kulveit et al., 2025): as AI capabilities improve and deployment expands, institutions may become less reliant on human cognition, weakening the mechanisms through which democratic systems remain responsive to citizens. Historically, this interdependence has operated across multiple, partially independent sources of social power. Mann (1986) distinguishes four such sources: economic, ideological, military, and political. General-purpose AI threatens to erode all four simultaneously by reducing citizen leverage in each domain. At the same time, the actors who control these systems may acquire increasing political influence as institutions become dependent on their infrastructure.

In the economic domain, if productivity gains flow disproportionately to capital owners, citizens lose bargaining power in ways the rentier-state literature predicts. This trajectory is consistent with the direction of technical change since the 1980s, in which automation has increasingly displaced labor rather than augmenting it (Acemoglu & Restrepo, 2019). In the ideological domain, AI introduces new mechanisms for centralizing belief formation. As discussed in Sections 3.1 and 3.2, reliance on a small number of similarly tuned models can narrow the range of publicly available framings, while personalized reinforcement loops can harden individual positions against revision. Beyond these passive effects, AI systems are increasingly capable of active persuasion, with evidence suggesting they can exceed human persuasive baselines (Rogiers et al., 2024; Hackenburg et al., 2025; Schoenegger et al., 2026). In the political domain, AI-powered surveillance and predictive enforcement enable more targeted suppression of political dissent at lower cost than traditional methods, reducing the state's dependence on broad cooperation to maintain order

(Xu, 2021; Beraja et al., 2023). In the military domain, autonomous systems remain largely developmental, but the trajectory suggests reduced reliance on human personnel for the exercise of coercive force, further weakening the bargaining power that conscription-based militaries historically afforded citizens (Simmons-Edler et al., 2024; Tilly, 1992).

Declining citizen leverage may shift institutional dependence away from citizens and toward the actors who control advanced AI systems. As frontier models become essential infrastructure for production, administration, and information access, control over the most capable systems, along with the compute and expertise required to deploy them, may concentrate in a small number of individuals, companies and states (Hendrycks et al., 2023; Bengio, 2023). Even open-weight models depend on complex supply chains for compute, data, and infrastructure that remain concentrated (Meiklejohn et al., 2025; Casper et al., 2025). Institutions may then become less dependent on citizens and more dependent on model providers, whose control over AI behavior and access gives them growing political influence.

What distinguishes the current moment from previous episodes of power concentration is not only the magnitude or speed of any single shift, but also their simultaneity. Historically, concentration in one domain often left others available as channels for countervailing pressure. When economic elites captured political institutions, ideological mobilization and the threat of unrest could restore balance (Acemoglu & Robinson, 2006). When states expanded coercive capacity, economic indispensability and normative delegitimation constrained overreach (Tilly, 1992; Acemoglu & Robinson, 2019). General-purpose AI is distinctive because it can reduce citizen leverage across economic, ideological, political and military domains simultaneously, narrowing the space for the cross-domain compensation that has historically made concentration correctable. These domains are also not independent: economic resources can be converted into political influence, political surveillance can suppress the organizing that might otherwise check economic concentration, and ideological control can prevent citizens from recognizing the erosion. The result is a potential for self-reinforcing concentration in which erosion in one domain accelerates erosion in the others.

T7: As AI capabilities grow and deployment expands, democratic governance may erode through the simultaneous weakening of citizen leverage across multiple domains of social power. By substituting for human labor, cognition, and participation across economic, ideological, political, and military domains, AI systems can weaken the sources of citizen leverage that historically sustained democratic accountability. At the same time, control over advanced models, compute infrastructure, and deployment pipelines may concentrate in a small number of companies and states, shifting institutional dependence away from citizens and toward AI providers. Together, these dynamics create the potential for self-reinforcing power concentration. These dynamics intensify as AI systems become more capable and more widely embedded in economic production and governance processes.

4. Recommendations

The failure modes we describe operate through different mechanisms (scale and volume, training incentives, architectural choices, institutional gaps) and no single intervention addresses all of them. We organize our recommendations around four complementary directions: simulating institutional risks before they materialize, training models to support rather than degrade epistemic health, constraining AI architecture in governance contexts, and building the institutional infrastructure that makes the other three enforceable.

4.1. Stress-test institutions through simulation

R1: Develop multi-agent simulations to evaluate institutional resilience under AI-mediated participation. The failure modes we identify require evaluation methods that go beyond model-level benchmarks to assess how individually benign outputs aggregate into systemic institutional effects (Yadav et al., 2025b; Pandey et al., 2025). Agent-based modeling has a long history in computational social science for studying how micro-level behaviors produce macro-level institutional outcomes (Epstein & Axtell, 1996). Recent work demonstrates that LLM-powered agents can replicate human behavior in social experiments (Park et al., 2023; Argyle et al., 2023; Aher et al., 2023) and simulate legislative processes (Baker & Azher, 2024). We propose extending these methods to institutional stress-testing: simulating how public comment systems, regulatory pipelines, and deliberative forums behave under large-scale AI-generated participation. Concrete questions include: at what volume does a comment system's signal-to-noise ratio collapse? How does opinion diversity in a simulated public change when a majority of participants use the same foundation

model? How does the cost of correcting a false belief scale with the duration of sycophantic reinforcement? These are empirical questions that can be studied in simulation before they play out in practice. While LLM-based simulation inherits biases from the models it uses (a methodological limitation that must be acknowledged and calibrated against real-world baselines), it offers a tractable path toward the population-level and institutional-scale evaluations that current AI safety frameworks lack (Costello et al., 2024; Salvi et al., 2025). Simulation-based stress-testing should be complemented by continuous monitoring of deployed systems: periodic retesting as AI capabilities evolve and usage patterns change, and cross-organizational sharing of incidents and near misses to enable early detection of emerging systemic risks.

4.2. Train models to support epistemic health

R2: Develop alignment methods that go beyond harm avoidance to instill pro-social epistemic behavior. Current alignment methods are primarily defensive: they train models to avoid harmful outputs through refusal, filtering, and reward shaping (Ouyang et al., 2022; Bai et al., 2022a). This is necessary but insufficient for the sociopolitical risks we describe, which arise not from models doing harmful things but from models doing helpful things in ways that degrade the systems around them (T1, T2; Sections 3.1, 3.2). A complementary agenda would train models to actively support epistemic health: to seek genuine understanding rather than agreement, to flag uncertainty rather than confabulate, and to value productive disagreement over user satisfaction. Existing work on truthfulness, calibration, and sycophancy reduction has begun pursuing some of these properties at the model level (Lin et al., 2022; Kadavath et al., 2022; Evans et al., 2021; Sharma et al., 2024), but frames them primarily as attributes of individual outputs rather than as properties of the model's role within a broader epistemic system. This distinction matters because the properties that matter most for sociopolitical resilience, such as cooperation, honest disagreement, and epistemic humility, are inherently relational: they describe how a model behaves toward other agents, not just the accuracy of its individual outputs. As work on cooperative AI has argued, cooperation problems are fundamentally multi-agent and cannot be reduced to single-system optimization (Dafoe et al., 2021). Recent evidence shows that advanced LLMs frequently fail to sustain cooperation in strategic settings, defaulting to free-riding, defection, or other socially harmful strategies (Piedrahita et al., 2025b; Cobben et al., 2026; Tewolde et al., 2026). The same logic extends to epistemic behavior: single-agent RLHF can optimize for truthfulness in isolation, but it cannot teach a model when to push back, when to defer, or how to navigate genuine disagreement productively. Multi-agent training environments where sycophancy leads to worse collective outcomes and honest pushback leads to better ones offer a more direct path (Irving et al., 2018). This research direction builds on existing work in cooperative AI (Dafoe et al., 2021), pluralistic alignment (Sorensen et al., 2024), and the philosophical case for navigating competing moral frameworks under reasonable pluralism (Gabriel, 2020). Concretely, public AI systems should be architected to support running multiple models in parallel, expose common interfaces for reasoning traces and decision logs, and enable systematic comparison of outputs on identical inputs. Cross-model disagreement checks and periodic re-benchmarking against alternative models can surface blind spots that arise when similar training produces similar biases (Gabriel & Keeling, 2025).

4.3. Constrain AI architecture in governance contexts

R3: For governance-adjacent AI, limit autonomy to preserve accountability and reduce systemic risk. The failure modes we describe generally intensify when AI systems act autonomously: agentic systems can flood participatory channels without human involvement (T3; Section 3.3), run private reinforcement loops that accumulate over months (T2; Section 3.2), make thousands of consequential decisions before anyone evaluates the normative commitments embedded in their design (T5, T6; Sections 3.5, 3.6), and accelerate the substitution of human roles on which democratic leverage depends (T7; Section 3.7). A growing body of work converges on the principle that combining high autonomy, broad generality, and superhuman capability in a single system creates qualitatively different risks than any two of these properties alone (Aguirre, 2023; Bengio et al., 2025). Drexler (2019) argues that general intelligence can emerge from specialized, composable services rather than monolithic agents, and recent autonomy taxonomies demonstrate that capability need not entail autonomy: designers can choose lower autonomy levels even for highly capable systems (Morris et al., 2023). For governance-adjacent AI (systems that touch democratic infrastructure, public administration, or institutional decision-making), this raises a question that deserves systematic investigation: under what conditions does increasing AI autonomy degrade institutional accountability faster than it improves institutional performance? The tool AI tradition suggests that capable, even superhuman systems can be designed to operate without autonomous goal-pursuit (Aguirre, 2023; Bengio et al., 2025), and the failure modes in this paper suggest that such constraints would reduce exposure to several sociopolitical risks simultaneously. Whether tool-like deployment is sufficient, or whether intermediate autonomy levels can be made safe with adequate oversight infrastructure, is an open empirical question, but one that should be answered before governance-adjacent systems are deployed at high autonomy by default. This principle should be encoded in Institutional

Safety Levels (R4.4) and procurement standards (R4.4), so that the degree of autonomy permitted scales with the strength of accountability mechanisms in place. We note that economic pressures may push toward greater autonomy over time (Branwen, 2016), making institutional constraints on deployment form, not just model behavior, an important complement to technical alignment.

4.4. Build institutional infrastructure

R4: Establish capability-triggered governance thresholds for public-sector AI. Each major institution (legislatures, courts, regulatory agencies, electoral systems) should formalize threat models that identify how AI alters their input, processing, and feedback layers, and specify capability thresholds at which risks emerge and cascade.[3] We propose encoding these thresholds as Institutional Safety Levels (ISLs) for public-sector AI deployment, analogous to the capability-triggered frameworks that frontier AI developers have adopted for model-level risks (Anthropic, 2023; Google DeepMind, 2024; OpenAI, 2025d). Each ISL binds concrete AI capabilities to mandatory procedural safeguards. For example, in a court system, the shift from "AI drafts internal research memos" to "AI generates sentencing or bail recommendations" would automatically trigger disclosure to affected parties, retention of full reasoning traces, and mandatory human sign-off with appeal pathways. Higher-impact uses, such as drafting legally operative text or enabling population-scale filing, would require an external audit or pre-deployment authorization. The thresholds themselves should be set through processes that include democratic input, not determined unilaterally by technical experts, and should be validated through the adversarial simulations described in Section 4.1 (Vaintrob, 2025).

R5: Require decision records and participation authentication for institutional AI. Institutional AI systems should log decision records by default: durable, standardized traces that capture inputs, model and prompt versions, tool calls, retrieved sources, intermediate state, and uncertainty, in formats suitable for audit, comparison, and legal review (Mitchell et al., 2019; Raji et al., 2020). These records should be queryable across cases so decisions can be audited and stress-tested, building on emerging standards for AI system logging (International Organization for Standardization, 2025; UK Government Digital Service, 2024). Explanations need to move beyond chain-of-thought, which can be unreliable (Arcuschin et al., 2025), toward methods that let officials test how outputs change when inputs change, whether through improving the faithfulness of reasoning traces themselves (Yu et al., 2025) or through causal and counterfactual explanation methods that make input-output dependencies explicit (Wachter et al., 2018). Separately, deployed systems need to track provenance and support proof-of-personhood for inputs such as public comments, filings, reports, or petitions, so institutions can distinguish genuine participation from automated volume without compromising privacy. Recent work on personhood credentials proposes privacy-preserving approaches using verifiable credentials and zero-knowledge proofs that avoid the centralization risks of biometric systems (Adler et al., 2024). Together, these measures make AI-mediated governance inspectable, contestable, and reliable at scale.

R6: Require interoperability and multi-provider strategies in public AI procurement. When a single model family is deployed across public institutions, its training data, safety filters, and reward functions effectively standardize how arguments are framed and which claims are treated as legitimate, producing epistemic monoculture (T1, T2; Sections 3.1, 3.2) and normative capture (T6; Section 3.6). Public procurement frameworks should require transparency regarding model capabilities, safety constraints, and update policies; mandate interoperability standards and data portability so that institutions are not locked into a single provider; and support multi-provider deployment strategies that enable hot-swapping without re-engineering workflows. As recent analyses of AI procurement argue, procurement is increasingly the de facto site of AI governance: consequential normative decisions are made through bilateral vendor negotiations without public accountability (Carr-Ryan Center for Human Rights Policy, 2026). Decision-log retention and explainability should be mandatory procurement requirements, modeled after communication and data retention standards in regulated sectors such as finance, so that model behavior, disagreements, and normative trade-offs remain auditable over time.

R7: Invest in deliberative infrastructure for AI governance. The failure modes we describe are partly artifacts of institutional designs (open comment systems, broadcast-model communication, unilateral standard-setting) that predate AI. Complementing defensive measures (R4.1–R4.4) with *proactive institutional innovation* can make governance channels more resilient by design. Concretely: fund structured deliberation tools (e.g., bridging-based platforms that surface consensus rather than amplify volume (Small et al., 2021)), establish standing citizens' assemblies with formal advisory roles and respond-or-explain requirements (i.e., regulators must either adopt assembly recommendations or publicly justify departing from them) in AI regulation, and pilot democratic input processes for consequential AI decisions such as model constitutions and safety thresholds (Huang et al., 2024; Weyl et al., 2024). Recent work demonstrates that AI-mediated deliberation can operate at

[3] Reliably measuring AI capabilities remains an open challenge, so ISLs should not depend on technical measurement alone; they should also incorporate public input on where thresholds belong and err toward requiring the procedural safeguards described in Section 4.4 when measurement is uncertain.

scale: Tessler et al. (2024) showed that an LLM-based mediator generated group statements preferred over those from human mediators across thousands of participants, while OpenAI's Democratic Inputs to AI program funded ten teams each engaging 500+ participants (OpenAI, 2023). A 2024 case demonstrates the approach for AI-specific harms: when deepfake advertisements impersonating public figures proliferated on social media, Taiwan convened 447 randomly selected citizens in 44 virtual deliberation rooms; an AI dialogue engine synthesized their proposals the same day. The assembly converged on actor-and-behavior regulation (platform liability for unsolicited deepfakes, mandatory labeling of unsigned ads, and throttling of non-compliant services) rather than content moderation. The law passed with multiparty support, and impersonation ads fell by 94% within a year (Siddarth et al., 2024; Reuters Investigates, 2025). These channels, designed around structured deliberation and identity-verified random selection, are inherently more robust to congestion and manipulation (T3, T4; Sections 3.3, 3.4) than open-submission systems, provided they include ongoing feedback from the communities affected by their outcomes.

5. Alternative Perspectives

Societies will gradually adapt without intervention. From a Hayekian perspective, complex social systems reach equilibrium through decentralized self-adaptation rather than centralized design (Hayek, 2013; Scott, 2020). Some argue that AI-induced disruptions will be absorbed through evolving social norms, market incentives, and trust heuristics (Folke et al., 2005), and that proactive intervention may overestimate our ability to anticipate and manage such systems. Historical precedent offers some support, as institutions adapted to the printing press and internet by developing new procedures and oversight mechanisms over time (Wu, 2010; Kissinger et al., 2021). However, the timeline for institutional adjustment may be compressed from decades to years, and the adaptation this view assumes may be structurally undermined by the dynamics it expects to resolve. Acemoglu (2021) argues that AI development concentrating economic rents also erodes the political and organizational capacity on which institutional self-correction depends. If this reinforcing dynamic holds, harms may accumulate faster than self-adaptation can resolve them, not only because the pace of change is unprecedented but because the capacity for institutional response is itself degraded.

Sufficient alignment will prevent sociopolitical risks. This view holds that advances in technical alignment will mitigate sociopolitical AI risks at the model level, reducing the need to analyze institutional dynamics (Russell, 2019; Ouyang et al., 2022; Bai et al., 2022b). We agree that alignment is necessary and that it can directly help with specific sociopolitical risks: less sycophantic training reduces belief reinforcement (Section 3.2), improved chain-of-thought faithfulness strengthens institutional auditability (Section 3.5), and models that refuse to generate deceptive content limit the most blatant epistemic attacks (Section 3.4). However, the relationship between alignment and sociopolitical risks is more complex than this view suggests. Some failure modes, particularly those driven by cost asymmetries and scale, persist regardless of alignment quality: a perfectly aligned model still makes it cheap to overwhelm a public comment system (Section 3.3), and alignment does nothing to prevent the displacement of human labor and participation that weakens citizen leverage over institutions (Section 3.7). Others are direct consequences of how current alignment methods work: the same reward signals that make models safe and helpful also narrow output diversity (Section 3.1) and embed developer values into public infrastructure (Section 3.6). Moreover, whether alignment delivers its sociopolitical benefits depends on institutional conditions: who sets the alignment objectives, how compliance is verified, and whether deployment contexts negate model-level gains. Alignment without functioning oversight institutions risks optimizing for the wrong objectives; institutional resilience without alignment lacks the technical tools to govern model behavior. The two agendas are mutually dependent.

6. The Way Forward

In this paper, we argue that sociopolitical risks from AI emerge at the level of institutions and governance systems, and therefore cannot be resolved by model-level alignment alone. Advancing this agenda requires coordinated action across multiple communities. For the AI research community, this means extending safety work toward system-level evaluations that capture aggregation effects, institutional load, and belief dynamics under realistic deployment conditions. For AI developers, it entails treating contestability, auditability, and pluralism as core design principles, not post-hoc remedy. For policymakers, it requires moving beyond reactive controls toward institution-level safeguards that preserve democratic responsiveness at scale. Crucially, institutional resilience need not be built from scratch: civic technology initiatives have demonstrated that structured deliberation and participatory governance can operate at a national scale—though adapting these tools to AI governance remains an open research challenge. We call on all three communities, together with the publics most affected by these systems, to ensure that rapidly advancing AI is matched by corresponding progress in institutional resilience. These agendas are not competing priorities: model-level alignment depends on functioning institutions to determine whose values are encoded and to verify compliance, while institutional resilience increasingly depends on tech-

nical tools to govern model behavior at scale. Advancing both in coordination is essential.

Acknowledgments

We thank Brenda Baker, Eric Grosse, Aidan Kierans, Stephan Schwahlen and Marcelo Sartori Locatelli for helpful feedback and discussions on the paper. This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by Coefficient Giving; and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

Acemoglu, D. Harms of AI. Working Paper 29247, National Bureau of Economic Research, 2021.

Acemoglu, D. and Restrepo, P. Automation and new tasks: How technology displaces and reinstates labor. *Journal of economic perspectives*, 33(2):3–30, 2019.

Acemoglu, D. and Robinson, J. *Economic Origins of Dictatorship and Democracy*. Cambridge University Press, 2006. URL https://EconPapers.repec.org/RePEc:cup:cbooks:9780521855266.

Acemoglu, D. and Robinson, J. A. *The Narrow Corridor: States, Societies, and the Fate of Liberty*. Penguin Press, New York, 2019. ISBN 978-0-735-22438-4.

Adler, S., Hitzig, Z., Jain, S., Brewer, C., Chang, W., DiResta, R., Lazzarin, E., McGregor, S., Seltzer, W., Siddarth, D., Soliman, N., South, T., Spelliscy, C., Sporny, M., Srivastava, V., Bailey, J., Christian, B., Critch, A., Falcon, R., Flanagan, H., Duffy, K. H., Ho, E., Leibowicz, C., Nadhamuni, S., Rozenshtein, A. Z., Schnurr, D., Shapiro, E., Strahm, L., Trask, A., Weinberg, Z. A. Y., Whitney, C. D., and Zick, T. Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online. *ArXiv*, abs/2408.07892, 2024.

Aguirre, A. Keep the future human. *arXiv preprint arXiv:2311.09452*, 2023.

Aher, G., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

AI Incident Database. Incident 348: Youtube recommendation reportedly pushed election fraud content to skeptics disproportionately. AI Incident Database, 2020. URL https://incidentdatabase.ai/cite/348/. Accessed: 2026-01-28.

AI Incident Database. Incident 198: Deepfake video of ukrainian president yielding to russia posted on ukrainian websites and social media. AI Incident Database, 2022. URL https://incidentdatabase.ai/cite/198/. Accessed: 2026-01-28.

Anthropic. Anthropic's responsible scaling policy, 2023. URL https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Version 3.0 released 2025.

Anthropic. Claude's constitution. Anthropic, 2025. URL https://www.anthropic.com/constitution. Accessed: 2026-01-28.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint*, 2025.

Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., Sorensen, T., and Wingate, D. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120 (41):e2311627120, October 2023. doi: 10.1073/pnas.2311627120. URL https://doi.org/10.1073/pnas.2311627120.

Atherton, D. Incident number 628: Fake Biden voice in robocall misleads New Hampshire Democratic voters in 2024 primary election. *AI Incident Database*, 2024. URL https://incidentdatabase.ai/cite/628. Accessed 2026-01-23.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. J., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022a.

This paper covers a wide range of empirical and normative topics, so with the exception of the corresponding author, inclusion as an author does not entail endorsement of all claims in the paper, nor does authorship imply an endorsement on the part of any individual's organization.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosiute, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b. URL https://arxiv.org/abs/2212.08073.

Baker, N. and Azher, J. Simulating the U.S. senate: An LLM-driven agent approach to modeling legislative behavior and bipartisanship. *arXiv preprint arXiv:2406.18702*, 2024.

Bakshy, E., Messing, S., and Adamic, L. A. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

Beblawi, H. and Luciani, G. (eds.). *The Rentier State*. Nation, State and Integration in the Arab World. Croom Helm, London, 1987.

Bengio, Y. Ai and catastrophic risk. *Journal of Democracy*, 34(4):111–121, 2023. doi: 10.1353/jod.2023.a907692. URL https://dx.doi.org/10.1353/jod.2023.a907692.

Bengio, Y., Cohen, M. K., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O. E., Rondeau, M.-A., St-Charles, P.-L., and Williams-King, D. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *ArXiv*, abs/2502.15657, 2025.

Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., Manning, S., Maslej, N., Mavroudis, V., McGlynn, C., Moulange, R., Newman, J., Ng, K. Y., Paskov, P., Rismani, S., Sastry, G., Seger, E., Singer, S., Stix, C., Velasco, L., Wheeler, N., Acemoglu, D., Conitzer, V., Dietterich, T. G., Heintz, F., Hinton, G., Jennings, N., Leavy, S., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Ramchurn, S. D., Russell, S., Schaake, M., Schölkopf, B., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Aguirre, L. A., Ajala, O., Albalawi, F., AlMalek, N., Busch, C., Collas, J., de Leon Ferreira de Carvalho, A. C. P., Gill, A., Hatip, A. H., Heikkilä, J., Johnson, C., Jolly, G., Katzir, Z., Kerema, M. N., Kitano, H., Krüger, A., Lee, K. M., Portillo, J. R. L., McLysaght, A., Molchanovskyi, O., Monti, A., Nemer, M., Oliver, N., Pezoa, R., Plonk, A., Ravindran, B., Riza, H., Rugege, C., Sheikh, H., Wong, D., Zeng, Y., Zhu, L., Privitera, D., and Mindermann, S. International ai safety report 2026, 2026. URL https://arxiv.org/abs/2602.21012.

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

Beraja, M., Kao, A., Yang, D. Y., and Yuchtman, N. Aitocracy*. *The Quarterly Journal of Economics*, 138(3): 1349–1402, 03 2023. ISSN 0033-5533. doi: 10.1093/qje/qjad012. URL https://doi.org/10.1093/qje/qjad012.

Bontridder, N. and Poullet, Y. The role of artificial intelligence in disinformation. *Data & Policy*, 3:e32, 2021. doi: 10.1017/dap.2021.20.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, United Kingdom, first edition edition, 2014. ISBN 9780199678112.

Branwen, G. Why tool AIs want to be agent AIs, 2016. URL https://gwern.net/tool-ai. Updated 2018. Accessed March 2026.

Buyl, M., Rogiers, A., Noels, S., Dominguez-Catena, I., Heiter, E., Romero, R., Johary, I., Mara, A. C., Lijffijt, J., and Bie, T. D. Large language models reflect the ideology of their creators. *Npj Artificial Intelligence*, 2(1):7, 2026.

Carlsmith, J. Is power-seeking AI an existential risk? *ArXiv*, abs/2206.13353, 2022.

Carr-Ryan Center for Human Rights Policy. Governance by procurement: How AI rights became a bilateral negotiation, 2026. URL https://www.hks.harvard.edu/centers/carr-ryan/our-work/carr-ryan-commentary/governance-procurement-how-ai-rights-became.

Casper, S., O'Brien, K., Longpre, S., Seger, E., Klyman, K., Bommasani, R., Nrusimha, A., Shumailov, I., Mindermann, S., Basart, S., Rudzicz, F., Pelrine, K., Ghosh, A., Strait, A., Kirk, R., Hendrycks, D., Henderson, P., Kolter, Z., Irving, G., Gal, Y., Bengio, Y., and Hadfield-Menell, D. Open technical problems in open-weight ai model risk management. *arXiv preprint*, 2025.

Castelló, I., Colleoni, E., Scherer, A. G., and Trittin-Ulbrich, H. Social media is a threat for democracy! a political perspective for analysing and diminishing harm. *Journal of Management Studies*,

n/a(n/a), 2025. doi: https://doi.org/10.1111/joms.70053. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/joms.70053.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Cobben, P., Huang, X. A., Pham, T. A., Dahlgren, I., Zhang, T. J., and Jin, Z. Gt-harmbench: Benchmarking ai safety risks through the lens of game theory, 2026. URL https://arxiv.org/abs/2602.12316.

Coeckelbergh, M. Democracy, epistemic agency, and ai: political epistemology in times of artificial intelligence. *AI and Ethics*, 3(4):1341–1350, 2023.

Costello, T. H., Pennycook, G., and Rand, D. G. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385, 2024.

Council of Europe. Council of europe framework convention on artificial intelligence and human rights, democracy and the rule of law, 2024. CETS No. 225.

Critch, A. and Russell, S. Tasra: a taxonomy and analysis of societal-scale risks from ai, 2023. URL https://arxiv.org/abs/2306.06924.

Dafoe, A. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., Hayes, J., Vyas, N., Merey, M. A., Brown-Cohen, J., Bunel, R., Balle, B., Cemgil, T., Ahmed, Z., Stacpoole, K., Shumailov, I., Baetu, C., Gowal, S., Hassabis, D., and Kohli, P. Scalable watermarking for identifying large language model outputs. *Nature*, 634:818 – 823, 2024.

Deutsch, K. W. *The Nerves of Government: Models of Political Communication and Control*. Free Press of Glencoe, London, 1963.

Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. Strong model collapse, 2024. URL https://arxiv.org/abs/2410.04840.

Drago, L. and Laine, R. The intelligence curse, 2025. URL https://intelligence-curse.ai/intelligence-curse.pdf. Accessed 2026-01-23.

Drexler, K. E. Reframing superintelligence: Comprehensive AI services as general intelligence. Technical Report 2019-1, Future of Humanity Institute, University of Oxford, 2019.

Durak, H. Y., Eğin, F., and Onan, A. A comparison of human-written versus AI-generated text in discussions at educational settings: Investigating features for ChatGPT, Gemini and BingAI. *European Journal of Education*, 2025.

Easton, D. *A Systems Analysis of Political Life*. John Wiley & Sons, New York, 1965.

Entman, R. M. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 02 2006. ISSN 0021-9916. doi: 10.1111/j.1460-2466.1993.tb01304.x. URL https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.

Epstein, J. M. and Axtell, R. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press and MIT Press, Washington, DC, 1996.

European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. OJ L, 2024/1689.

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., and Saunders, W. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

Farrell, H. and Newman, A. L. Weaponized interdependence: How global economic networks shape state coercion. *International Security*, 44:42–79, 2019.

Fiedler, A. and Döpke, J. Do humans identify AI-generated text better than machines? evidence based on excerpts from German theses. *International Review of Economics Education*, 2025.

Folke, C., Hahn, T., Olsson, P., and Norberg, J. Adaptive governance of social-ecological systems. *Annual Review Environment and Resources*, 30(1):441–473, 2005.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

Gabriel, I. and Keeling, G. A matter of principle? ai alignment as the fair treatment of claims. *Philosophical Studies*, 182(7):1951–1973, 2025. doi: 10.1007/s11098-025-02300-4.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomavsev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C.,

Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., Mateos-Garcia, J., Weidinger, L., Street, W., Lange, B., Ingerman, A., Lentz, A., Enger, R., Barakat, A., Krakovna, V., Siy, J. O., Kurth-Nelson, Z., McCroskery, A., Bolina, V., Law, H., Shanahan, M., Alberts, L., Balle, B., de Haas, S., Ibitoye, Y., Dafoe, A., Goldberg, B., Krier, S., Reese, A., Witherspoon, S., Hawkins, W., Rauh, M., Wallace, D., Franklin, M., Goldstein, J. A., Lehman, J., Klenk, M., Vallor, S., Biles, C., Morris, M. R., King, H., y Arcas, B. A., Isaac, W., and Manyika, J. The ethics of advanced AI assistants. *ArXiv*, abs/2404.16244, 2024.

Garrett, R. K. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285, 2009.

Gentzkow, M. and Shapiro, J. M. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

Gilens, M. and Page, B. I. Testing theories of american politics: Elites, interest groups, and average citizens. *Perspectives on politics*, 12(3):564–581, 2014.

Gillespie, T. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

Google DeepMind. Frontier safety framework, 2024. URL https://deepmind.google/blog/strengthening-our-frontier-safety-framework/. Version 3.0 released 2025.

Green, R. Foia-flooded elections. *Ohio State Law Journal*, 85:255–306, 2024.

Guess, A., Nyhan, B., Lyons, B., and Reifler, J. Avoiding the echo chamber about echo chambers. *Knight Foundation*, 2(1):1–25, 2018.

Habermas, J. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, 1996.

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H. Z., Rand, D. G., and Summerfield, C. The levers of political persuasion with conversational AI. *CoRR*, abs/2507.13919, 2025. doi: 10.48550/ARXIV.2507.13919. URL https://doi.org/10.48550/arXiv.2507.13919.

Hao, Q., Xu, F., Li, Y., and Evans, J. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649(8099):1237–1243, 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09922-y. URL https://www.nature.com/articles/s41586-025-09922-y.

Hardt, M. and Mendler-Dünner, C. Performative prediction: Past and future. *Statistical Science*, 40(3):417–436, 2025.

Hayek, F. A. *The fatal conceit: The errors of socialism*. Routledge, 2013.

Hays, K. and Jamali, L. Trump orders government to stop using anthropic in battle over ai use, 2026. URL https://www.bbc.com/news/articles/cn48jj3y8ezo. BBC News article, Accessed: 2026-02-28.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023. URL https://arxiv.org/abs/2306.12001.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., and Trautsch, A. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13, 2023.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., and Couzin, I. D. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19(1):46–54, 2015. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2014.10.004. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(14)00233-2.

Holbein, J. Left Behind? Citizen Responsiveness to Government Performance Information. *American Political Science Review*, 110(2):353–368, 2016. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055416000071. URL https://www.cambridge.org/core/journals/american-political-science-review/article/left-behind-citizen-responsiveness-to-government-performance-information/C387F8656931C3C877AA6B7A1CD1148B.

Howard, P. N. *Lie machines: How to save democracy from troll armies, deceitful robots, junk news operations, and political operatives*. Yale University Press, 2020.

Hsiao, Y.-T., Lin, S.-Y., Tang, A., Narayanan, D., and Sarahe, C. vTaiwan: An empirical study of open consultation process in Taiwan. *SocArXiv Preprints*, 2018. doi: 10.31235/osf.io/xyhft.

Hu, M. Cambridge analytica's black box. *Big Data & Society*, 7(2):1–6, 2020. doi: 10.1177/2053951720938091. URL https://doi.org/10.1177/2053951720938091.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. Collective constitutional AI: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil, 2024. Also available as arXiv:2406.07814.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., and Hardt, M. Algorithmic amplification of politics on twitter. *Proceedings of the national academy of sciences*, 119(1):e2025334119, 2022.

International Organization for Standardization. ISO/IEC DIS 24970: Artificial intelligence — AI system logging, 2025. Draft International Standard, under development.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Ischen, C. *Persuasive Agents: Unraveling the Persuasive Potential of Conversational Agents*. Carolin Ischen, 2022.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15, 2023.

Jin, Z., Heil, N., Liu, J., Dhuliawala, S., Qi, Y., Schölkopf, B., Mihalcea, R., and Sachan, M. Implicit personalization in language models: A systematic study. *arXiv*, abs/2405.14808, 2024.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022.

Karadal, P. and Kekulluoglu, D. Prioritize economy or climate action? investigating ChatGPT response differences based on inferred political orientation. *ArXiv*, abs/2511.04706, 2025.

Kissinger, H. A., Schmidt, E., and Huttenlocher, D. *The age of AI: and our human future*. Hachette UK, 2021.

Klapper, J. T. *The effects of mass communication.* Free press, 1960.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint arXiv:2501.16946*, 2025.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, X., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Bras, R. L., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. In *Proceedings of the 2nd Conference on Language Modeling (COLM 2025)*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH#discussion.

Landemore, H. Open democracy and digital technologies. In *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton University Press, 2020. URL https://pacscenter.stanford.edu/wp-content/uploads/2020/12/Chapte-2_9780226748436_1stPages_Mktg-2-1.pdf.

Lau, H. T. and Zubiaga, A. Understanding the effects of human-written paraphrases in LLM-generated text detection. *Natural Language Processing Journal*, 11:100151, 2025.

Layne, N. Insight: Pro-Trump activists swamp election officials with sprawling records requests. *Reuters*, August 2022.

Lazarsfeld, P. F., Berelson, B., and Gaudet, H. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press, 1968.

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

Levin, R. M. The duty to respond to rulemaking comments. *Yale Law Journal Forum*, 134:821, 2024.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022.

Mann, M. *The Sources of Social Power: Volume 1, A History of Power from the Beginning to AD 1760*. Cambridge University Press, Cambridge, 1986. ISBN 978-0-521-31349-0.

Mansbridge, J. Everyday talk in the deliberative system. *Deliberative politics*, pp. 211–239, 1999.

March, J. G. Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1):71–87, 1991. ISSN 1047-7039. URL https://www.jstor.org/stable/2634940.

McCombs, M. E. and Shaw, D. L. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2): 176–187, 1972.

Meiklejohn, S., Blauzvern, H., Maruseac, M., Schrock, S., Simon, L., and Shumailov, I. Position: Machine learning models have a supply chain problem. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=zfohnbkMu0.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 220–229. ACM, 2019.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Levels of AGI: Operationalizing progress on the path to AGI. *arXiv preprint arXiv:2311.02462*, 2023. Published at ICML 2024.

Motoki, F., Pinho Neto, V., and Rodrigues, V. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1):3–23, 2024.

National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0), 2023.

OpenAI. Democratic inputs to AI, 2023. URL https://openai.com/index/democratic-inputs-to-ai/.

OpenAI. Expanding on what we missed with sycophancy, 2025a. URL https://openai.com/index/expanding-on-sycophancy/. OpenAI post. Accessed 2026-01-23.

OpenAI. Sycophancy in GPT-4o: What happened and what we're doing about it, 2025b. URL https://openai.com/index/sycophancy-in-gpt-4o/. OpenAI post. Accessed 2026-01-23.

OpenAI. Openai model spec (version 2025-12-18). OpenAI Model Specification, 2025c. URL https://model-spec.openai.com/2025-12-18.html. Accessed: 2026-01-28.

OpenAI. Our updated preparedness framework, 2025d. URL https://openai.com/index/updating-our-preparedness-framework/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022.

Padmakumar, V. and He, H. Does writing with language models reduce content diversity? In *International Conference on Learning Representations*, 2024.

Pandey, P. S., Le, H. S., Bhardwaj, D., Mihalcea, R., and Jin, Z. Socialharmbench: Revealing llm vulnerabilities to socially harmful requests, 2025. URL https://arxiv.org/abs/2510.04891.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, New York, NY, USA, 2023. Association for Computing Machinery.

Pasquale, F. The black box society: The secret algorithms that control money and information. In *The black box society*. Harvard University Press, 2015.

Piedrahita, D. G., Strauss, I., Schölkopf, B., Mihalcea, R., and Jin, Z. Democratic or authoritarian? probing a new dimension of political biases in large language models, 2025a. URL https://arxiv.org/abs/2506.12758.

Piedrahita, D. G., Yang, Y., Sachan, M., Ramponi, G., Schölkopf, B., and Jin, Z. Corrupted by reasoning: Reasoning language models become free-riders in public goods games, 2025b. URL https://arxiv.org/abs/2506.23276.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 33–44. ACM, 2020.

Rathi, I. M., Taylor, S., Bergen, B., and Jones, C. GPT-4 is judged more human than humans in displaced and inverted turing tests. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pp. 96–110. International Conference on Computational Linguistics, 2025.

Reuters Investigates. Meta created a playbook to fend off pressure to crack down on scammers, documents show. *Reuters*, December 2025. Available at https://www.reuters.com/investigations/meta-created-playbook-fend-off-pressure-

crack-down-scammers-documents-show-2025-12-31/.

Rogiers, A., Noels, S., Buyl, M., and Bie, T. D. Persuasion with large language models: a survey. *CoRR*, abs/2411.06837, 2024. doi: 10.48550/ARXIV.2411.06837. URL https://doi.org/10.48550/arXiv.2411.06837.

Rosenthal, R. W. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2 (1):65–67, 1973.

Ross, M. L. Does oil hinder democracy? *World Politics*, 53 (3):325–361, 2001. doi: 10.1353/wp.2001.0011.

Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, 2019. ISBN 978-0525558613.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-generated text be reliably detected? *ArXiv*, abs/2303.11156, 2023.

Salvi, F., Horta Ribeiro, M., Gallotti, R., and West, R. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, pp. 1–9, 2025.

Sandbrink, J. B. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*, 2023. URL https://arxiv.org/abs/2306.13952.

Schmidt, V. A. Democracy and Legitimacy in the European Union Revisited: Input, Output and 'Throughput'. *Political Studies*, 61(1):2–22, 2013. ISSN 0032-3217. doi: 10.1111/j.1467-9248.2012.00962.x. URL https://doi.org/10.1111/j.1467-9248.2012.00962.x.

Schoenegger, P., Salvi, F., Liu, J., Nan, X., Debnath, R., Fasolo, B., Leivada, E., Recchia, G., Günther, F., Zarifhonarvar, A., Kwon, J., Islam, Z. U., Dehnert, M., Lee, D. Y. H., Reinecke, M. G., Kamper, D. G., Kobaş, M., Sandford, A., Kgomo, J., Hewitt, L., Kapoor, S., Oktar, K., Kucuk, E. E., Feng, B., Jones, C. R., Gainsburg, I., Olschewski, S., Heinzelmann, N., Cruz, F., Tappin, B. M., Ma, T., Park, P. S., Onyonka, R., Hjorth, A., Slattery, P., Zeng, Q., Finke, L., Grossmann, I., Salatiello, A., and Karger, E. When large language models are more persuasivethan incentivized humans, and why, 2026. URL https://arxiv.org/abs/2505.09662.

Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Ressa, M., Schweitzer, F., Song, D., Summerfield, C., Tang, A., Bavel, J. J. V., van der Linden, S., and Kunst, J. R. How malicious ai swarms can threaten democracy. *Science*, 391(6783):354–357, 2026. doi: 10.1126/science.adz1697. URL https://www.science.org/doi/abs/10.1126/science.adz1697.

Scott, J. C. *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press, 2020.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Shapira, I., Benade, G., and Procaccia, A. D. How RLHF Amplifies Sycophancy, 2026. URL http://arxiv.org/abs/2602.01002.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., DURMUS, E., Hatfield-Dodds, Z., Johnston, S., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y. URL https://www.nature.com/articles/s41586-024-07566-y.

Shur-Ofry, M. Multiplicity as an AI governance principle. *Indiana Law Journal*, 100(4):1527–1564, June 2025. URL https://www.repository.law.indiana.edu/ilj/vol100/iss4/6/.

Siddarth, D., Huang, S., and Tang, A. A vision of democratic AI. *The Digitalist Papers*, 1, 2024. Available at https://www.digitalistpapers.com/essays/a-vision-of-democratic-ai.

Simmons-Edler, R., Badman, R., Longpre, S., and Rajan, K. Ai-powered autonomous weapons risk geopolitical instability and threaten ai research. *arXiv preprint arXiv:2405.01859*, 2024.

Small, C., Bjorkegren, M., Erkkila, T., Shaw, L., and Megill, C. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: Revista de Pensament i Anàlisi*, 26(2), 2021. doi: 10.6035/recerca.5516.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment, 2024.

Stephenson, M. C. A Costly Signaling Theory of "Hard Look" Judicial Review. *Administrative Law Review*, 58 (4):753–813, 2006. ISSN 0001-8368. URL https://www.jstor.org/stable/40711887.

Strange, S. *The retreat of the state: The diffusion of power in the world economy*. Cambridge University Press, 1996.

Summerfield, C., Argyle, L. P., Bakker, M., Collins, T., Durmus, E., Eloundou, T., Gabriel, I., Ganguli, D., Hackenburg, K., Hadfield, G. K., Hewitt, L., Huang, S., Landemore, H., Marchal, N., Ovadya, A., Procaccia, A., Risse, M., Schneier, B., Seger, E., Siddarth, D., Skaug Sætra, H., Tessler, M. H., and Botvinick, M. The impact of advanced AI systems on democracy. *Nat Hum Behav*, 9 (12):2420–2430, October 2025.

Swenson, A. and Weissert, W. New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary. AP News, Jan 2024. URL https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5. Accessed 2026-01-23.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M. M., and Summerfield, C. Ai can help humans find common ground in democratic deliberation. *Science*, 386, 2024.

Tewolde, E., Zhang, X., Piedrahita, D. G., Conitzer, V., and Jin, Z. Coopeval: Benchmarking cooperation-sustaining mechanisms and llm agents in social dilemmas, 2026.

Tilly, C. *Coercion, Capital, and European States, AD 990–1992*. Studies in Social Discontinuity. Blackwell, Cambridge, MA, revised edition, 1992. ISBN 978-1-557-86368-3.

Tischer, J. F. Panmemic inoculation: How Taiwan is nerfing the pandemic with cute humour. *East Asian Journal of Popular Culture*, 8(2):183–204, 2022. doi: 10.1386/eapc_00073_1.

Turner, C. and Eisikovits, N. Programmed to Please: The Moral and Epistemic Harms of AI Sycophancy, 2026. URL https://papers.ssrn.com/abstract=6117867.

UK Government Digital Service. Algorithmic transparency recording standard hub, 2024. URL https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub. Mandatory for central government departments since February 2024.

Vaintrob, L. The AI adoption gap: Preparing the US government for advanced AI. Forethought Research, 2025. URL https://www.forethought.org/research/the-ai-adoption-gap. Accessed: 2026-01-28.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018.

Waller, I. and Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature*, 600 (7888):264–268, 2021.

Weaver, R. K. The Politics of Blame Avoidance. *Journal of Public Policy*, 6(4):371–398, 1986. ISSN 0143-814X. URL https://www.jstor.org/stable/4007281.

Weber, M. *Economy and Society: An Outline of Interpretive Sociology*, volume 2. University of California Press, 1978.

Weidinger, L., Mellor, J. F. J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S. M., Hawkins, W. T., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W. S., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359, 2021.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL https://dl.acm.org/doi/10.1145/3531146.3533088.

Weyl, E. G., Tang, A., and the Plurality Community. *Plurality: The Future of Collaborative Technology and Democracy*. Plurality Community, 2024. ISBN 9798869327116. CC0 public domain. Available at https://plurality.net/.

Woolley, S. C. and Howard, P. N. *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018.

Wu, F., Black, E., and Chandrasekaran, V. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025a. URL https://openreview.net/forum?id=yZ7sn9pyqb.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., and Wong, D. F. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51:275–338, 2025b.

Wu, T. *The Master Switch: The Rise and Fall of Information Empires*. Alfred A. Knopf, 2010.

Xu, X. To repress or to co-opt? authoritarian control in the age of digital surveillance. *American Journal of Political Science*, 65(2):309–325, 2021. doi: https://doi.org/10.1111/ajps.12514. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12514.

Yadav, N., Liu, J., Ortu, F., Ensafi, R., Jin, Z., and Mihalcea, R. Revealing hidden mechanisms of cross-country content moderation with natural language processing, 2025a. URL https://arxiv.org/abs/2503.05280.

Yadav, N., Ortu, F., Liu, J., Yook, J., Schölkopf, B., Mihalcea, R., Cazzaniga, A., and Jin, Z. Are llms good safety agents or a propaganda engine?, 2025b. URL https://arxiv.org/abs/2511.23174.

Yang, Z., Zhao, G., and Wu, H. Watermarking for large language models: A survey. *Mathematics*, 13(9), 2025.

Yu, X., Wang, Z., Yang, L., Li, H., Liu, A., Xue, X., Wang, J., and Yang, M. Causal sufficiency and necessity improves chain-of-thought reasoning. *arXiv preprint*, abs/2506.09853, 2025. URL https://arxiv.org/abs/2506.09853. arXiv:2506.09853 [cs.CL].

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *ArXiv*, abs/2504.13837, 2025.

Zaleznik, D. Facebook and genocide: How facebook contributed to genocide in myanmar and why it will not be held accountable. Research paper, Systemic Justice Journal, Critical Corporate Theory Collection, Harvard Law School, 2021. URL https://systemicjustice.org/wp-content/uploads/2021/10/Zaleznik_FinalPaper.pdf. Accessed: 2026-02-28.

Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, 2019. ISBN 9781610395694.