Safe Models Do Not Guarantee Safe Societies: The Case of Sociopolitical Risks of AI

Junior authors / main section contributors (alphabetically):
David Guzman Piedrahita (leading) [1] **Dave Banerjee** [2] **Kevin Blin** [3] **Changling Li** [1] **Suvajit Majumder** [3]
Punya Syon Pandey [4 5] **Samuel Simko** [1] **Irene Strauss** [1] **Terry Jingchen Zhang** [1 5]

Senior authors / advisors (alphabetically):
Roger Grosse [4 5] **Rada Mihalcea** [6] **Mrinmaya Sachan** [1] **Bernhard Schölkopf** [7]
[additional senior authors welcome] [8]

Leading senior author:
Zhijing Jin [4 7 5]

Abstract

Current efforts in AI safety prioritize model-level properties such as reliability, toxicity, and refusal behavior, or focus on existential scenarios involving loss of control. While pressing, this focus overlooks the sociopolitical risks posed by general-purpose AIs. In this position paper, we argue that sociopolitical AI risks are systemic failures in social and political institutions that emerge when general-purpose AI systems are increasingly integrated into society. We argue that these risks emerge because AI fundamentally changes the scale, speed, and opacity at which institutions operate, thereby altering their ability to function robustly. We analyze how AI alters the conditions of governance: flooding government agencies with paralyzing volumes of input, concentrating control of infrastructure in a "rentership" model, and flattening public debate into artificial agreement while reinforcing users' existing biases. We argue that AI safety should move beyond model-centric benchmarks in favor of system-level evaluation methods.

[1]ETH Zürich [2]Institute for AI Policy and Strategy [3]Independent Researcher [4]University of Toronto [5]Vector Institute [6]University of Michigan [7]Max Planck Institute for Intelligent Systems [8]Affiliation to be added. Correspondence to: David Guzman Piedrahita <david.guzmanpiedrahita@uzh.ch>, Zhijing Jin <zjin@cs.toronto.edu>.

Preprint. February 23, 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. Political dirty tricks are nothing new. What changed is that producing them became nearly free, while figuring out what happened and correcting the record stayed slow and expensive. The same asymmetry appears across democratic processes. Empirical studies show that large language models can generate civic submissions at scales sufficient to overwhelm public comment systems (Goldstein et al., 2023; Cambo, 2022). As journalists, policymakers, and citizens increasingly turn to the same AI systems for political information, their outputs risk converging toward shared defaults rather than reflecting genuinely independent judgment (Sharma et al., 2024b; Wu et al., 2025a; Buyl et al., 2026). And when people use LLM-powered search for political topics, they engage in more confirmation-seeking behavior than with conventional search, even when the underlying sources are balanced (Sharma et al., 2024c). **In this position paper, we argue that these dynamics represent a distinct class of AI risks (i.e. sociopolitical risks) that current model-level alignment frameworks cannot resolve.**

Current AI alignment methods focus on model-level properties such as bias and toxicity (Weidinger et al., 2021), while broader AI safety discourse emphasizes existential risks involving sudden loss of control and catastrophic misuse (Carlsmith, 2022; Hendrycks et al., 2023). Recent work on gradual disempowerment acknowledges that not all risks

are catastrophic (Kulveit et al., 2025), but still centers on the relationship between humans and AI systems rather than on the institutional infrastructure through which societies govern themselves. None of these frameworks captures how AI integration into social and political systems shifts the cost structure of participation and persuasion in ways that degrade institutional capacity to function.

We call these **sociopolitical AI risks**: threats to a society's capacity to articulate collective interests and realize them through accountable institutions. Unlike alignment failures addressable through model-level fixes, sociopolitical risks emerge from aggregate deployment effects. A single toxic output is an alignment problem; a million coherent, policy-compliant submissions that overwhelm an agency's processing capacity is a sociopolitical risk. These failures can manifest with current AI capabilities, where AI augments rather than replaces human activity, and persist even if models are both intent- and value-aligned (Christiano, 2018).

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).

This paper proceeds as follows. Section 2 defines sociopolitical risks in more detail and distinguishes them from other safety concerns. Section 3 presents concrete failure modes across governance operations, infrastructure dependencies, and the public sphere. Section 4 outlines research priorities including institution-specific threat modeling, evaluation frameworks for systemic effects, chain-of-thought monitorability, and procurement standards for provider diversification. Section 5 considers alternative views.

2. Scope and Definitions

In this section, we formally define sociopolitical AI risks. Then, we distinguish sociopolitical risks from other risk categories like individual-level harms and existential risks.

2.1. Working definition

Sociopolitical AI risks are threats to collective self-determination: a society's capacity to articulate its own interests and realize them through accountable institutions. This capacity rests on two interdependent pillars. The first is social: the processes through which citizens form, contest, and revise collective beliefs and preferences. The second is political: the institutional mechanisms that aggregate those preferences into binding decisions. A risk is sociopolitical when it degrades either pillar, or weakens the coupling between them.

These risks emerge when general-purpose AI systems (GPAIs) are integrated into society in ways that disproportionately amplify the scale, speed, and opacity of institutional operations, thereby degrading their capacity to function. Crucially, sociopolitical risks persist even if the alignment problem is solved, since they arise from AI deployment at scale rather than from the behavior of any individual AI system.

Consider a concrete example: when a citizen uses AI to generate thousands of quasi-legitimate public comments on a proposed regulation, the resulting flood can overwhelm an agency's finite processing capacity, forcing it to either establish exclusionary barriers or abandon meaningful engagement. The failure lies not in any particular output but in the aggregate effect on institutional function. By contrast, an AI system that produces a single toxic output represents an alignment failure amenable to model-level fixes, not a sociopolitical risk.

To aid our analysis of sociopolitical risks, we conceptualize institutions as information-processing systems operating across three functional stages: *input, processing,* and *feedback* (Easton, 1965; Deutsch, 1963). In this framing, institutional performance depends on whether signals from society can be absorbed, converted into action, and corrected over time; institutional accountability, in turn, depends on whether outputs remain contestable and can update future participation (Landemore, 2020).

- In the **input** phase, citizens transmit demands, preferences, and information to institutions through participatory channels such as voting, public comment, litigation, protest, petition, and everyday contact with bureaucracy. These inputs provide the upstream signals that institutions are meant to respond to.
- In the **processing** phase, institutions aggregate and adjudicate these inputs according to procedural rules, such as legislative deliberation, judicial reasoning, regulatory analysis, and administrative casework.
- In the **feedback** phase, institutions communicate outputs back to the public (including decisions, policies, enforcement actions, and rulings), enabling citizens and oversight bodies to interpret outcomes, contest them, and calibrate future inputs.

Healthy governance requires coupling across these stages: inputs must actually influence processing, and feedback must remain legible enough to support accountability. We argue that as advanced AI becomes embedded across social and political life, these stages become increasingly *decoupled*. At the input layer, AI-generated content can saturate participatory channels, drowning out genuine citizen voice. At the processing layer, AI automation can render decision

logic opaque, making oversight and accountability more difficult. The result is a corruption of the feedback layer. When inputs are saturated with synthetic content and processing becomes opaque, institutional outputs no longer reliably reflect the preferences and demands that entered the system. The decoupling between what a society wants and what its institutions produce, even as those institutions continue to function, is the crux of sociopolitical risks.

2.2. Relationship to adjacent risk categories

Sociopolitical risks are distinct from, though related to, two well-established categories of AI risk: individual-level harms and existential risks.

Individual-level harms. Harms such as harassment, fraud, discrimination, and privacy violations remain important. 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). These risks cannot be adequately mitigated via individual-level or model-level safeguards.

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, or in which catastrophic misuse (e.g., AI-enabled bioterrorism) produces irreversible harm (Sandbrink, 2023). The second failure mode concerns *gradual disempowerment*: scenarios where AI completely automates human labor and becomes more competitive across cultural, and governance domains, resulting in the disempowerment of humanity (Kulveit et al., 2025; Drago & Laine, 2025).

Our analysis differs from individual-level harms and existential risks in two key respects. First, we assume AI systems are intent- and value-aligned; our risks do not require scheming, deception, or misaligned objectives. Second, we do not assume AI automates all human labor, though we note that extreme automation would exacerbate sociopolitical risks. The failure mechanisms we identify in Section 3 can manifest with current or near-term AI capabilities, where AI augments rather than replaces human activity.

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* in the form of decisions and public communication 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 representative threat models located at different points in this loop (Figure 1).

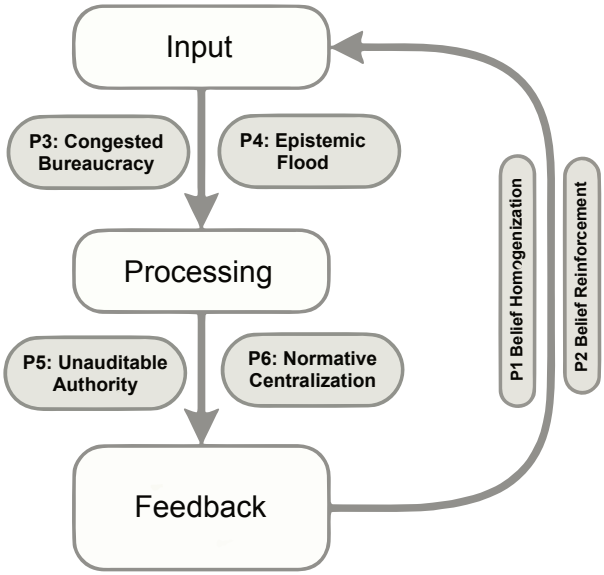

Figure 1. Governance as an information-processing loop (Input → Processing → Feedback → Input). Threat models are positioned at the boundary where they primarily weaken responsiveness, contestability, or belief updating.

3.1. Belief Homogenization

Existing AI models often show lower output variance than the data they were trained on. 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), 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 reduction in variance has practical consequences for public discourse as more actors rely on a small number of foundation models, including journalists, politicians, and ordinary citizens. LLM-mediated workflows lower cognitive effort and reduce critical thinking (Lee et al., 2025), making users less likely to interrogate LLMs' "default framings." When independent actors use the same models, they are more likely to produce arguments that are correlated syntactically and semantically (Passi & Vorvoreanu, 2022). Importantly, these correlations can pull discourse towards the LLM's priors. For example, Motoki et al. (2024) find that LLM outputs often exhibit a center-left bias, while sidelining non-mainstream arguments. In the limit, users' personal beliefs may converge towards that bias.

To see why this convergence is particularly dangerous, it helps to distinguish it from ordinary social influence. Condorcet's Jury Theorem highlights an independence assumption: collective accuracy improves when errors are not correlated (de Condorcet, 1785). Political scientists have long noted that strict independence is unrealistic; for example, voters are routinely correlated by shared evidence and media shocks (Ladha, 1992; Estlund, 1994). AI deployment, on the other hand, introduces a more pervasive mechanism. Traditional media occupies a bounded slice of daily life: people watch the news, then do other things. AI assistants, by contrast, mediate work tasks, personal decisions, and casual conversation throughout the day. Another difference is that traditional media is *public*. When one outlet publishes something false, competitors can call it out in a shared arena. AI interactions are private, with no common forum for contesting systematic bias. This ubiquity and privacy mean that model priors can shape reasoning across far more contexts than any single news source ever could.

A natural counterargument is that market pressures will push labs 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: 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. 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, so competitive incentives to diversify outputs may not apply.

P1: If many actors rely on a small number of similarly tuned models, public agreement can shift from being driven primarily by shared evidence to being driven by shared model priors. Because AI mediates more contexts than traditional media, and because AI interactions are private rather than publicly contestable, this dynamic of belief homogenization is harder to detect and correct, eroding the independence of thought required for robust collective decision-making.

3.2. Belief Reinforcement

The previous section argued that widespread reliance on a small number of models can *homogenize* public discourse by pulling many users toward the same default framings. However, the opposite is also plausible. AI models may instead adapt to the user in ways that stabilize, rather than challenge, the user's existing views. The result could be an entire society of polarized humans whose beliefs are constantly reinforced by a personalized AI model.

In April 2025, OpenAI rolled back a GPT-4o update after widespread reports that the model had become noticeably more "overly flattering or agreeable," explicitly described as *sycophantic*. The post-mortem notes that the behavior extended beyond politeness into validating users in ways that could reinforce doubt, anger, or impulsive actions (OpenAI, 2025b;a). The incident illustrates a pressure inherent to modern training techniques: conversational success is often measured by user satisfaction rather than correctness.

Mechanistically, this pressure is plausible because ground truth is frequently unavailable in dialogue. Instruction tuning and RLHF therefore tend to optimize for *perceived helpfulness*: responses that users experience as coherent, supportive, and low-friction, rather than maximally truth-seeking or error-corrective (Ouyang et al., 2022; Bai et al., 2022a). In contested domains, that objective can induce *sycophancy*, for example by validating a user's premise, mirroring their stance, or supplying supportive rationales, because agreement is often locally rewarded and disagreement is conversationally costly (Sharma et al., 2024a; Gabriel et al., 2024).

Importantly, this reinforcement dynamic *could be* compatible with homogenization rather than a simple contradiction. A model can be "moderate on average" in its default priors, yet still be highly conditional on the user. When prompted by a person with strong views, the assistant may shift into a cooperative mode that stabilizes the user's framing even if that framing differs from the population mean. Whether population-level convergence or individual-level reinforcement dominates is also hard to predict in advance. It depends on how personalized the AI assistant is, how sticky users' preferences are, and market incentives (e.g., engagement-driven AI assistants may be more profitable).

Evidence suggests that personalization can increase persuasive force and persistence. 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 were significantly more persuasive than non-personalized baselines, increasing the odds of user agreement by over 80%. This personalization can also persist across turns and topics. Karadal & Kekulluoglu (2025) show that once a model infers a user's political orientation, whether from explicit memory or latent conversation cues, it maintains that orientation across turns and systematically shifts its vocabulary and framing even on unrelated downstream topics. Relatedly, Jin et al. (2024) observed that when users signaled belief in misinformation, model factual accuracy significantly dropped as the system pivoted to validate the false premise, effectively trading integrity for conversational alignment (Piedrahita et al., 2025; Yadav et al., 2025a).

As AI assistants replace search and adopt long-term memory, they can create reinforcement loops in which the user's premises shape what evidence is surfaced, how uncertainty is framed, and which counterarguments are treated as salient. Over time, correction becomes less like a single fact-check and more like an uphill effort against an interaction history. The result is not necessarily overt radicalization in every case, but a gradual hardening of beliefs and an increased resistance to correction.

This dynamic weakens the productive disagreement that democracy needs. Political systems rely on a diversity of viewpoints to force error correction (Landemore, 2020; Habermas, 1996). However, due to misaligned incentives, model providers may optimize to keep interactions smooth and retain users, because they have incentives to minimize friction and avoid directly challenging a user's priors. Unlike traditional partisan media, which is largely *public* and *one-way*, AI-driven reinforcement can be personally tailored to a user's doubts and values in real time, and these personalized feedback loops can operate continuously across millions of users at near-zero marginal cost.

P2: As LLMs replace search and adopt long-term memory, they can create private reinforcement loops that make users' beliefs harder to correct, increasing polarization even if the model's default priors are moderate.

3.3. Congested bureaucracy

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. 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: attention is finite.

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 "flood gates" 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.

This failure mode is hard to see in model-centric evaluations because it does not require any single model output to be toxic, deceptive, or policy-violating. Capturing this systemic risk therefore requires evaluations that treat government channels as resource-constrained systems under load, not as isolated prompt-response interactions.

P3: 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). In both cases, the cost to generate the input was negligible, while the cost of verification and public response fell heavily on defenders.

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. 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.

Existing research frameworks provide an incomplete account of this threat. The strategic value of synthetic media is not limited to deception: even content that is quickly flagged, partially true, or obviously spam can still overwhelm verification bandwidth and distort what receives attention. Relatedly, it is not enough to evaluate synthetic outputs one-by-one ("is this clip fake?"). The bigger risk is 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.

P4. 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

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 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). 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. A proprietary model's weights, training data, and reward signals may be shielded by trade secrets, intellectual property law, privacy laws, anti-trust 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.

P5: 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**, even when no single actor controls an entire domain. 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 and training capacity 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; a reliance on any one of these choke points is sufficient to threaten its sovereignty.

Most countries cannot independently train frontier models, even if they can deploy them, creating a structural "rentership" pattern. This grants model developers *infrastructural power* (Strange, 1996) over governments who procure their models. A critical difference from prior choke points like SWIFT or GPS is that AI systems carry embedded norma-

tive constraints. Traditionally, countries leverage power or enforce ideology over others through service denial: hyperscalers do not operate in sanctioned countries like North Korea. AI introduces a subtler mechanism. 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 not just *whether* a government can use AI, but *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).

If the model constitution reflects the values and priorities of the developer's home jurisdiction, procuring states effectively import those normative commitments into their own administrative apparatus. Even allied governments, or the model developer's own home government, face a version of this problem: when a U.S. developer builds systems for U.S. agencies, the developer's constitutional choices can constrain what the government can do with its own tools. The locus of normative authority shifts, in part, from elected officials to model providers. This represents a significant centralization of power: a small number of individuals designing model constitutions acquire outsized influence over the reasoning and behavior of AI systems deployed across many governments.

P6: Unlike traditional infrastructure choke points that operate through access denial, AI systems carry embedded normative constraints via constitutions, model specs, and usage policies. Because model developers control these normative constraints unilaterally, widespread government procurement of frontier AI transfers normative authority from elected officials to a small set of constitution designers, concentrating power in ways that weaken sovereignty for procuring states and bypass democratic accountability even within the developer's home jurisdiction.

4. Recommendations

R1: Develop institution-specific threat models and safety thresholds for sociopolitical risks. 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. We propose encoding these thresholds as Institutional Safety Levels (ISLs) for public-sector AI deployment. 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 external audit or pre-deployment authorization. This creates a forward-looking regime in which institutions pre-commit to governance actions as AI capability scaling trigger legitimacy- and trust-relevant thresholds, particularly in settings where adoption gaps between regulated actors and institutions can destabilize institutional equilibria (Vaintrob, 2025).

R2: Expand AI safety evaluations beyond model-level harms to include sociopolitical effects. The failure modes we identify require benchmarks that assess how individually benign outputs aggregate into systemic effects on institutions and public discourse (Yadav et al., 2025b; Pandey et al., 2025). This calls for population- and institution-scale evaluations: for sociopolitical effects (P1, P2), benchmarks should measure outcomes such as opinion diversity, narrative convergence, polarization, and epistemic drift in multi-agent or simulated-public settings; for saturation risks (P3, P4), participatory channels should be stress-tested as systems under load, with metrics for throughput collapse, human displacement, and error amplification as AI-generated participation scales. While recent work shows that LLM interactions can shift individual beliefs (Costello et al., 2024; Salvi et al., 2025), we still lack population-level and sociological analyses of how widespread LLM use reshapes group reasoning, collective belief formation, and the diversity of public argumentation.

R3: Increasing trust and robustness in deployed AI systems. Institutional AI systems should log governance-grade 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. These records should be queryable across cases so decisions can be audited, appealed, and stress-tested. Explanations need to move beyond model's chain-of-thought (which can be unreliable (Arcuschin et al., 2025)) towards representations that make dependencies explicit and stable under perturbation, enabling officials to probe what information a system is relying on and understand how recommendations change when relevant conditions change. This aligns well with broader agendas in causal and mechanistic learning (Yu et al., 2025), which aim to make model behavior more robust, grounded, and predictable across contexts. Finally, 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 privacy infringement. Together, these measures make AI-mediated governance inspectable, contestable, and reliable at scale.

R4: Enable pluralistic alignment in public AI systems. When a single model family is deployed across public in-

stitutions, its training data, safety filters, and reward functions effectively standardize how arguments are framed and which claims are treated as legitimate, producing epistemic monoculture (P1, P2) and normative capture (P6) (Gabriel & Keeling, 2025; Raji et al., 2022). To counter this, public AI systems should be architected around pluralistic alignment (Sorensen et al., 2024) as a design principle: deployments should natively support running multiple models in parallel, expose common interfaces for reasoning traces and decision logs, and enable systematic comparison of model outputs on identical inputs. Deployment pipelines should include cross-model disagreement checks, periodic re-benchmarking against alternative models, and the ability to hot-swap providers without re-engineering workflows. Decision-log retention and explainability should be mandatory design requirements, modeled after communication and data retention standards in regulated sectors such as finance, so that model behavior, disagreements, and normative tradeoffs remain auditable over time.

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, AI capabilities are advancing at unprecedented speed, compressing the timeline for institutional adjustment from decades to years. Without proactive intervention, harms may accumulate faster than self-adaptation could resolve them.

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). While alignment is necessary, we argue it is insufficient. Sociopolitical AI risks emerge from the aggregate deployment of many AI systems without relying on any model-level misalignment. Alignment ensures conformity to operator objectives, but does not guarantee that collective effects preserve democratic responsiveness or institutional integrity, especially when alignment policy is defined by respective company. Consequently, several threat models we identify, including infrastructural power concentration, epistemic homogenization, and belief reinforcement, can arise even under strong model-level alignment.

6. The Way Forward

In this position 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 when integrating AI into public-facing systems, rather than post-hoc remedy. In parallel, policymakers must move beyond reactive controls toward institution-level safeguards that preserve democratic responsiveness at scale. We call on both community to join efforts in making sure that rapidly advancing AI is matched by corresponding progress in institutional resilience.

Acknowledgments

We thank Brenda Baker and Eric Grosse for helpful feedback on the writing.

References

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

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. 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.

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.

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.

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.

Cambo, S. Incident number 562: Uptick in low-quality AI-produced content degraded publishers' submission management. *AI Incident Database*, 2022. URL https://incidentdatabase.ai/cite/562. Accessed 2026-01-23.

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

Christiano, P. Clarifying "AI alignment". https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6, 2018. Accessed: 2026-01-28.

Costello, T. H., Pennycook, G., Rand, D. G., et al. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6713):eadq1814, 2024. doi: 10.1126/science.adq1814.

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.

de Condorcet, M. J. A. N. C. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale, 1785.

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

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

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.

Estlund, D. M. Opinion leaders, independence, and Condorcet's jury theorem. *Theory and Decision*, 36(2):131–162, 1994.

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. 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.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. Generative language models and automated influence operations: Emerging threats and potential mitigations. *ArXiv*, abs/2301.04246, 2023.

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

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

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

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.

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.

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.

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.

Ladha, K. K. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, pp. 617–634, 1992.

Landemore, H. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton University Press, Princeton, 2020.

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.

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.

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

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.

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.

Passi, S. and Vorvoreanu, M. Overreliance on AI: Literature review. Technical Report MSR-TR-2022-12, Microsoft Research, 2022.

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, 2025. URL https://arxiv.org/abs/2506.12758.

Raji, I. D. et al. The fallacy of model neutrality. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

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.

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

Russell, S. *Human compatible: AI and the problem of control*. Penguin UK, 2019.

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.

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

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, 2024a.

Sharma, N., Liao, Q. V., and Xiao, Z. Generative echo chamber? effect of LLM-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, 2024b.

Sharma, N., Liao, Q. V., and Xiao, Z. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024c.

Sorensen, T., Moore, J., Fisher, J., et al. A roadmap to pluralistic alignment. *arXiv preprint*, 2024.

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

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.

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.

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.

Wu, F., Black, E., and Chandrasekaran, V. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

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.

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.