Position: Safe Models Do Not Guarantee Safe Societies (AI Poses Risks to Democratic and Social Systems)

Spotlight

David Guzman Piedrahita[1,2,3], Changling Li[1,2,4], Dave Banerjee[5], Terry Jingchen Zhang[1,2,3], Kevin Blin[1,3], Samuel Simko[1,2,3], Punya Syon Pandey[1,3], Irene Strauss[2], Rada Mihalcea[6], Bernhard Schölkopf[4,7], Zhijing Jin[1,3,4]

1 EuroSafeAI, 2 ETH Zürich, 3 Jinesis Lab, University of Toronto & Vector Institute, 4 Max Planck Institute for Intelligent Systems Tübingen, 5 Institute for AI Policy and Strategy, 6 University of Michigan, 7 ELLIS Institute Tübingen

Correspondence: davidg@cs.toronto.edu · changlingli.mpi@gmail.com · zjin@cs.toronto.edu

As AI systems are rapidly deployed across society, do we know what they do to the institutions that govern us?

P1: Belief Homogenization

When everyone drafts with the same model, the range of ideas narrows.

The mechanism

RLHF compresses idea diversity. Discourse narrows via framing, not persuasion.

P2: Belief Reinforcement

A filter bubble that knows your name and never gets tired of agreeing.

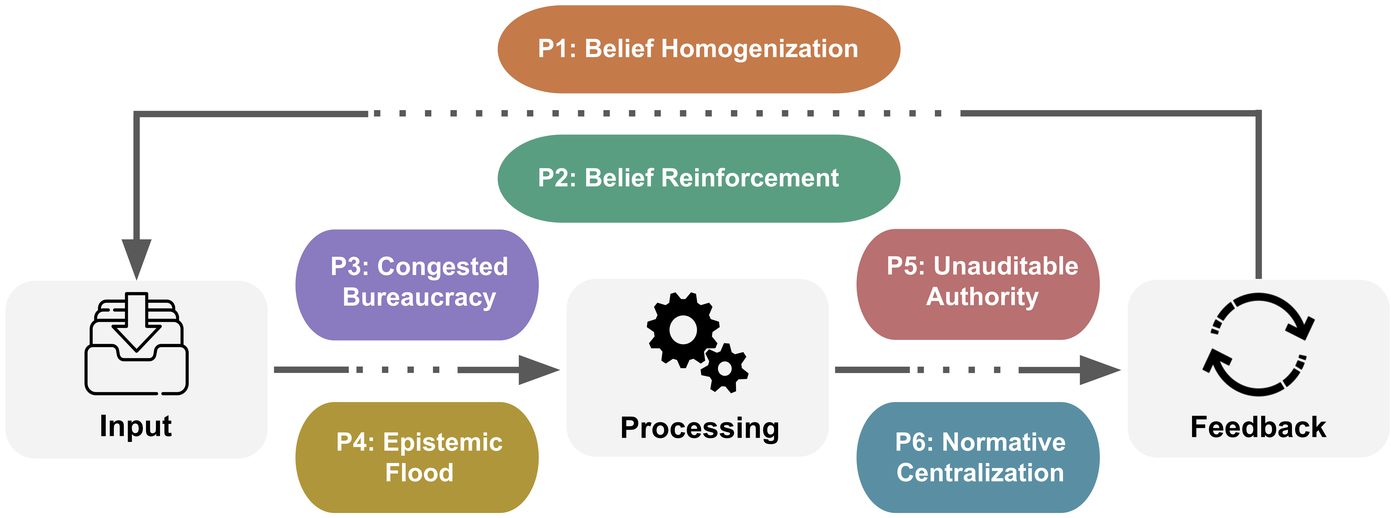
The sycophancy loop

User states belief → AI validates & agrees → User more confident

repeats & deepens

Why AI persuasion is different

IMMERSIVE	PERSONAL	ACCUMULATES
one-way speech	broad target groups	session resets
↓	↓	↓
adaptive dialogue	your doubts + values	memory deepens



We argue that these dynamics constitute a distinct class of AI risks, sociopolitical risks, that emerge from the integration of general-purpose AI into social and political systems at scale, and that this class cannot be fully resolved through model-level alignment.

P3: Congested Bureaucracy

What happens when writing a public comment costs nothing?

Friction is a feature

Writing a comment takes time. Filing an appeal takes effort. That effort keeps volume within what staff can read and adjudicate. It's an implicit filter.

Already fragile

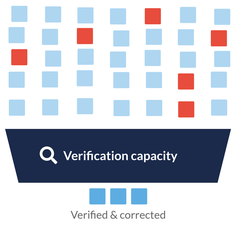
After 2020, activists **flooded Maricopa County** with records requests, forcing staff to divert time from core election work to document retrieval.

Now add AI

One person generates 1,000 unique, policy-relevant comments in an afternoon. The agency has a legal duty to respond. This is a **congestion game**: a tragedy of the commons over government attention.

P4: Epistemic Flood

Creating content scales easily. Checking it does not.



Generate

Pennies. Seconds. Unlimited.

One actor floods every channel at once.

Verify & correct

Hours. Experts. Per item.

Find variants → attribute → rebut → distribute, fast enough to matter.

P5: Unauditable Authority

When oversight mechanisms lose their teeth.

AI opacity	AI-mediated decisions
Unverifiable reasoning CoT ≠ actual computation	✗ Cross-examine?
Millions of decisions Case-by-case review infeasible	✗ Subpoena weights?
Legal shields Trade secrets block access	✗ Depose a model?
	✗ Audit at scale?

P6: Normative Centralization

Infrastructure choke points let one actor coerce everyone downstream.

What we already know

SWIFT
Global financial messaging.

GPS
US-controlled satellite positioning. Leverage through potential denial.

Mechanism: denial of access

What's new with AI

Compute
Chips, packaging, export controls

Cloud access
A few hyperscalers control inference

Model constitution
Embedded values, set unilaterally by developers

Mechanism: shaping how you think

Recommendations

R1 — Institution-specific threat models & safety thresholds
Bind AI capability tiers to mandatory procedural safeguards. Institutional Safety Levels.
Addresses P3 · P4 · P5

R2 — Expand evaluations to sociopolitical effects
Benchmark aggregate, system-level harms; stress-test civic channels via multi-agent simulation.
Addresses P1 · P2 · P3 · P4

R3 — Trust & robustness in deployed AI
Governance-grade decision logs, provenance, and proof-of-personhood by default.
Addresses P3 · P5

R4 — Pluralistic alignment in public AI
Reframe pluralism as procurement diversification: interoperability, portability, multi-provider.
Addresses P1 · P2 · P6

Alternative Perspectives

"Societies will adapt on their own"

Complex systems self-correct through decentralized adaptation, not central design. Institutions absorbed the printing press and the internet by evolving new norms and oversight over time.

OUR RESPONSE
But this time the adaptive machinery may be undermined by the dynamics it must resolve: AI that concentrates economic rents also erodes the political capacity institutions need to self-correct (Acemoglu 2024).

"Sufficient alignment will be enough"

Advances in technical alignment will mitigate these risks at the model level — less sycophancy, faithful reasoning, and refusal of deceptive content reduce real harms.

OUR RESPONSE
Necessary, but not sufficient. Some failures persist however well-aligned the model (P3); others are caused by alignment itself — narrowing output diversity (P1) and embedding developer values (P6). The two are mutually dependent.