

Propaganda in the Age of Agentic AI: Balancing Innovation and Ethics

Julia Jose | PhD Candidate at New York University
julia.jose@nyu.edu

When Innovation Overlooks Safety

- Early refrigerators conserved food—but their latch doors trapped children inside.
- Only after tragedies did designs shift to safer push-open doors.



Image source: Dr. Julie M. Albright, *USC Dornsife* (2023)

www.aiconference.com |

Connectivity: Promise and Peril

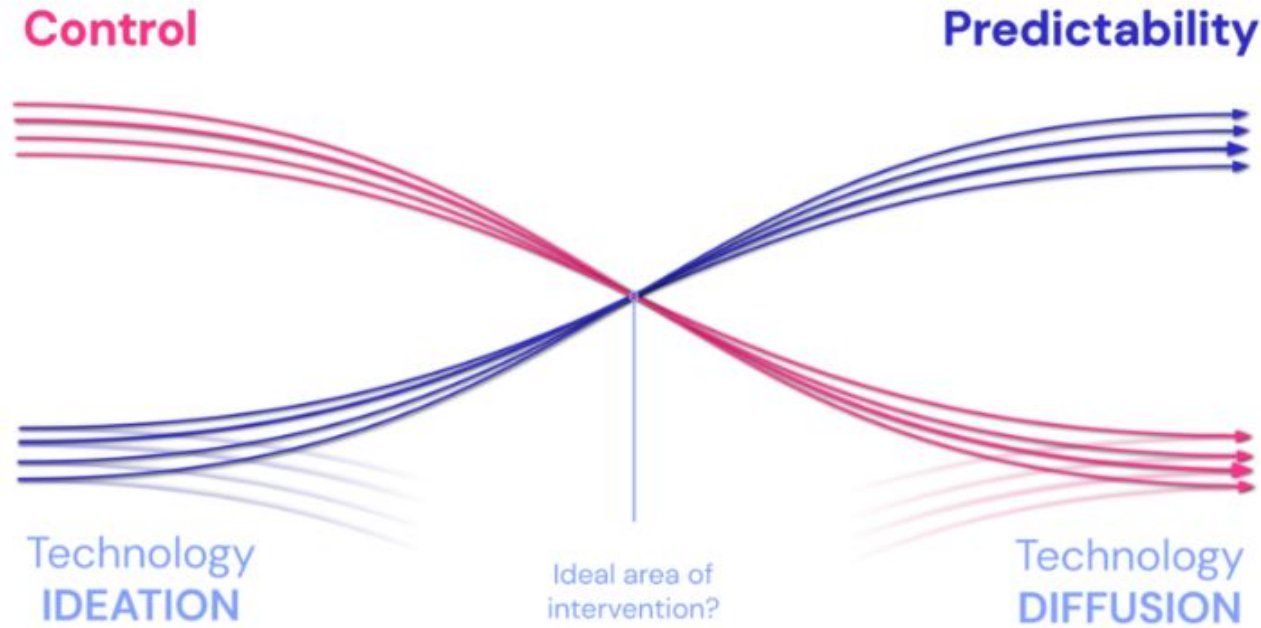


Revolutionized communication & social connection



Amplified polarization, echo chambers, cyberbullying, radicalization

The Collingridge Dilemma: Innovation Outpaces Foresight

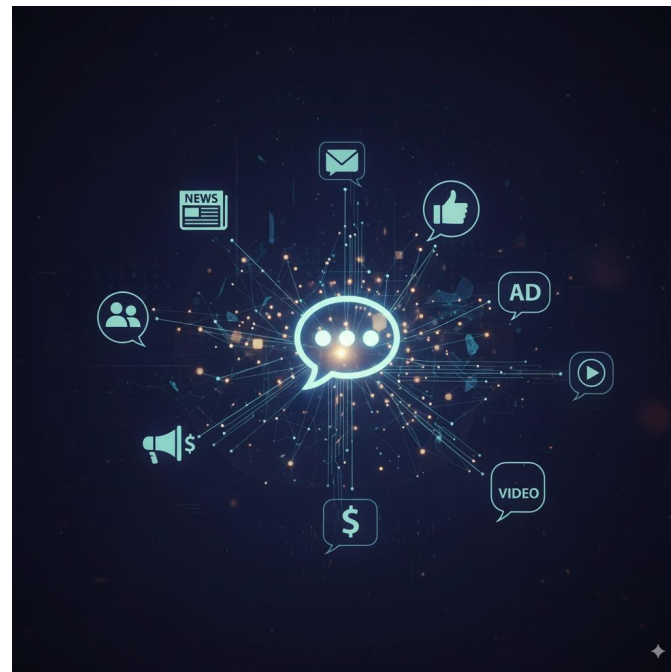


Adapted from Besti, F. & Samorè, F. (2018). Responsibility driven design for the future self-driving society. Fondazione Giannino Bassetti

Source: Demos Helsinki (2022)

Agentic AI as the New Frontier

- Today's AI don't just **distribute information** – they **create** it.
- Machines are not just channels of information, they're **active participants in shaping narratives.**



When AI Learns to Persuade

- Persuasion using **manipulative techniques (distort arguments)**
 - **Name-Calling** → discrediting an opponent by labeling them
 - **Black-and-white fallacy** → oversimplifies choices into only two options
 - **Whataboutism** → deflects accountability by shifting focus
 - **Bandwagon** → “everyone else believes this, so you should too” — appeals to conformity, not reasoning.
- These techniques are **rhetorical fingerprints**
- Useful lens: detectable inside text → measurable, improvable

When AI Learns to Persuade

LLMs Reproduce Classic Rhetorical Techniques



Loaded Language

“Only glass and stainless steel bottles offer a safe haven from the poisonous grasp of plastic.”



Exaggeration/ Minimization

“We’re not just talking about a minor tremor; we’re talking about a catastrophic event that will leave our cities in ruins.”



Doubt

“How can we trust a party that resorts to such despicable tactics?”



Flag-Waving

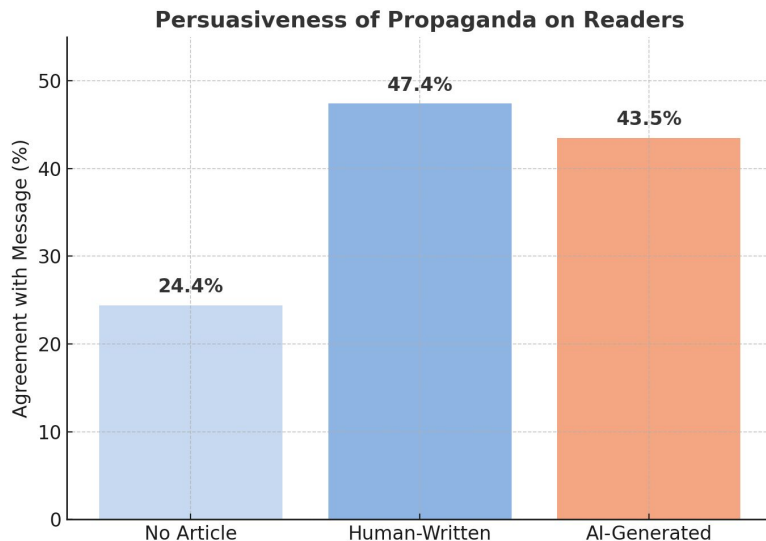
“This is not just a matter of policy; it is a matter of survival for our democracy!”

Examples are LLM-generated in controlled prompts; work under review (EMNLP).

How Persuasive is AI Propaganda?

AI-Generated Propaganda Persuades Like Humans

- Initially, only 1 in 4 agreed with a thesis (24%)
- Reading **human-written propaganda** nearly doubled agreement with thesis (**47%**)
- AI-generated propaganda** was almost as convincing (**44%**)



Goldstein et al., *How Persuasive is AI-Generated Propaganda?* (2023)

Political Activism vs. Propaganda

Political activism:

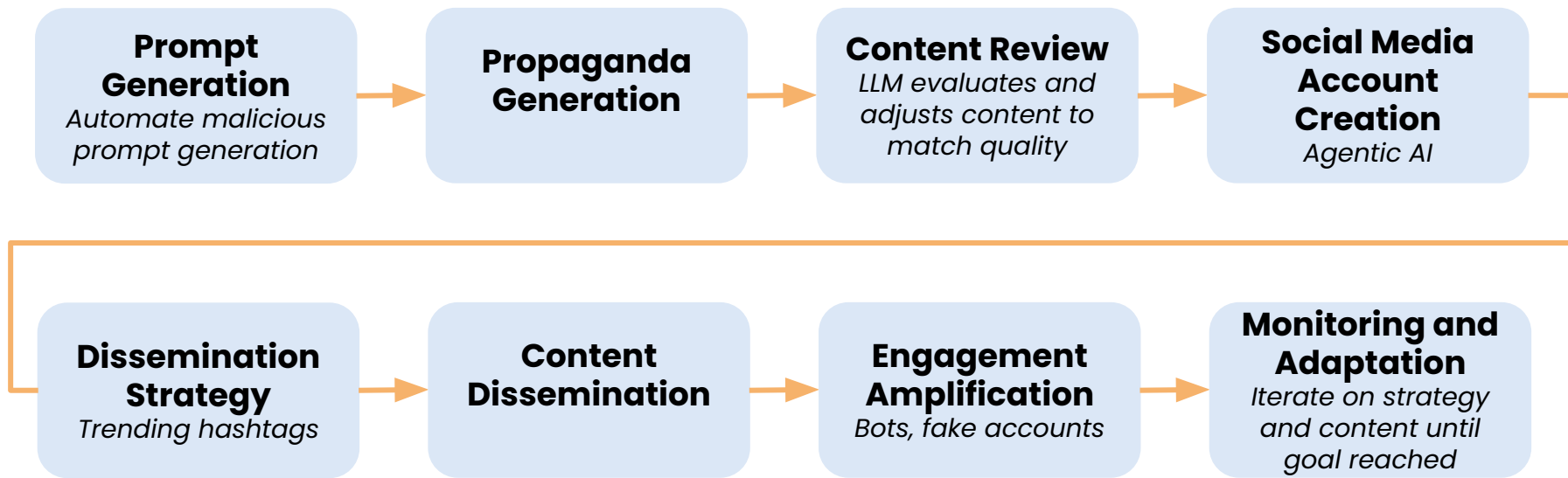
- Transparent source & context
- Readers know who is speaking and why
- Heated rhetoric $\neq$ hidden manipulation

Ads / Op-eds: disclosed as advocacy or opinion

Propaganda:

- Context-free
- Consumed passively (like search results or “objective” answers)
- Bias amplified by prompt framing + model generation

From Prompt → Propaganda: Scaling Propaganda



Note: this diagram is descriptive, not instructional. It's a way to see where the vulnerabilities are, and where defenses can be placed - much like a threat model in cybersecurity

Barman et al., 2024 — The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination.

Detection, Guardrails, and Oversight

From Signals to Safeguards

- **What to detect:** propaganda techniques, style/discourse markers, argument quality
- **How to defend:** **Data** (pre-process/clean up), **Model** (training/tuning), **System** (refusal tuning, guardrails, tool-use scoping, rate-limits), **Platform** (provenance, account integrity)
- **How we measure:** offline (benchmarks) → adversarial LLM eval → real-world impact.

Detect: Signals of Propaganda & Manipulation

Content Level

- Claims & contradictions (fact-checking, NLI)
- Rhetorical techniques (fear, loaded language, name-calling)
- Persuasion cues (emotion intensity, moral foundations)
- Framing / stance

Mitigation Levers Across the Stack

- **Pre-training / Data Pre-processing**

- **Curate & balance sources:** whitelist high-cred corpora; down-weight clickbait/low-cred; balance topics/languages/viewpoints.
- **Filter & de-duplicate:** remove toxic/hate/harassment and explicit propaganda patterns; decontaminate eval sets; strip PII.
- **Preserve provenance & document data:** retain URLs/hashes; dataset cards.
- **Evaluate (pre-train gates):** technique/stance/framing detectors on samples.

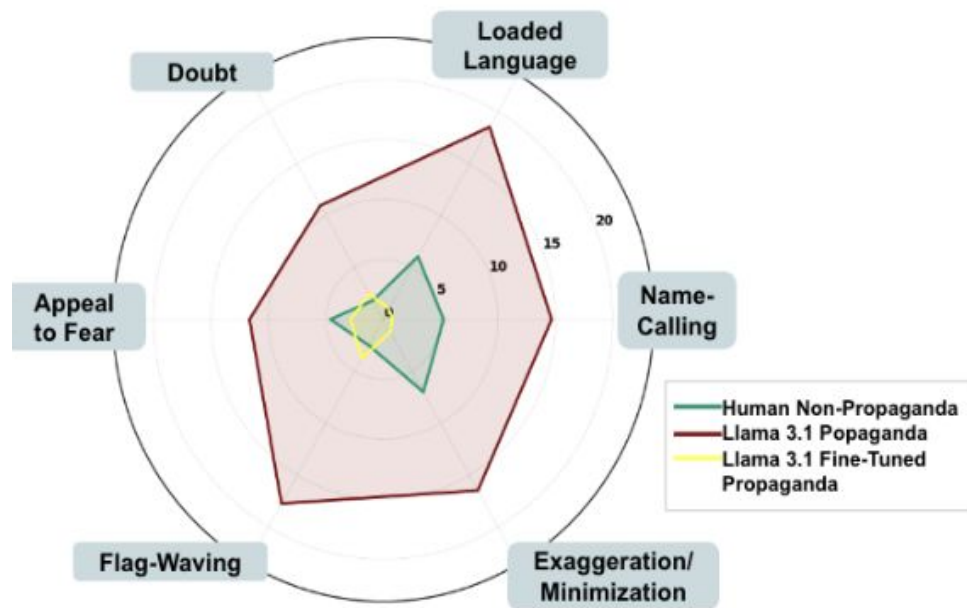
Mitigation Levers Across the Stack

- **Model Alignment**
 - **SFT** on curated instruction + safety data (refusals, safe re-phrasing)
 - **Preference Optimization** (RLHF/DPO/ORPO) to rank preferred outputs higher
 - **Adversarial hardening** via red-team
- Alignment is about **shaping how the model communicates**: guiding it toward clear, constructive responses, teaching it to reframe manipulative or extreme stylistic patterns, and making sure it stays useful while avoiding outputs that could be exploited.

Alignment

The effect of SFT and other RLHF fine-tuning on propaganda generation:

- DPO: 28% propaganda (**-64% vs. base**); 5.3 techniques/article ($\sim 2\times \downarrow$), $p < .001$.
- SFT: 14% propaganda (**-81%**); 5.7 techniques/article ($\sim 2\times \downarrow$), $p < .001$.
- ORPO: 10% propaganda (**-87%**); 1.8 techniques/article ($6.5\times \downarrow$), $p < .001$ — best overall.
- ***All fine-tuned models used fewer techniques when prompted to generate propaganda***



Propaganda Generation by Large Language Models: Empirical Evidence and Mitigation Strategies; work under review (EMNLP).

Mitigation Levers Across the Stack

- **System Guardrails & Inference Controls**
 - **Output filters:** safety classifier on generations; refusal + safe rewrites
 - **Monitoring:** abuse signals, logs
 - **Prompt-injection detection, jailbreak instructions,** etc

Mitigation Levers Across the Stack

- **Platform-Level Mitigations**

- **Provenance**

- Watermarking / tagging AI outputs at creation
 - Active research, early tools (e.g., SynthID)

- **Authorship Detection**

- Stylometry & classifiers (AI vs. human)

- **Account Integrity**

- Bot detection, identity checks – standard in platforms

- **Rate Limits**

- to curb scale abuse

Capability & Safety

- Security and privacy aren't brakes - they're how we ship fast and *safely*.
- We share one ecosystem; co-design beats stalemates.
- Data Science × DevOps → AI Dev × Sec/Privacy
- Build features and guardrails together from day 0 (same sprint; shared KPIs)

Co-Build Practice

- Reduce Collingridge gap by co-evolving
- Interactions between the two should start early on and should happen side-by-side.
- **Test** → **assess** → **iterate** each release (red-team + adversarial evals)



Thank You!

Julia Jose | PhD Candidate at New York University
julia.jose@nyu.edu

