Adverse Selection in the AI Data Commons

Evidence from News and Media Sites

Kai Zhu

Bocconi University

Digital Economy Workshop 2026

The AI Training Data Problem

- Generative AI derives its power from training on **high-quality web content**
  (Gunasekar et al. 2023; Longpre et al. 2023)

The AI Training Data Problem

- Generative AI derives its power from training on **high-quality web content**
  (Gunasekar et al. 2023; Longpre et al. 2023)
- Yet AI companies use this content **without compensation**
  Rare exceptions:
  - OpenAI–Reddit $60M deal
  - OpenAI–Associated Press licensing agreement

The AI Training Data Problem

- Generative AI derives its power from training on **high-quality web content**
  (Gunasekar et al. 2023; Longpre et al. 2023)
- Yet AI companies use this content **without compensation**
  Rare exceptions:
  - OpenAI–Reddit $60M deal
  - OpenAI–Associated Press licensing agreement
- No established market: producers' only option is **binary opt-out** via `robots.txt`
  (Kim et al. 2025; Chang & He 2025)

The AI Training Data Problem

- Generative AI derives its power from training on **high-quality web content**
  (Gunasekar et al. 2023; Longpre et al. 2023)
- Yet AI companies use this content **without compensation**
  Rare exceptions:
  - OpenAI–Reddit $60M deal
  - OpenAI–Associated Press licensing agreement
- No established market: producers' only option is **binary opt-out** via `robots.txt`
  (Kim et al. 2025; Chang & He 2025)

Who opts out? If the highest-quality producers exit first, the result is a **lemons problem** (Akerlof, 1970)—and the remaining AI data commons systematically degrades.

Research Questions & Preview of Findings

RQ: Does the market for AI training data exhibit adverse selection?

Key findings:

1. **Quality–blocking gradient**
   High-factual media block at **6**× the rate of low-factual sources
2. **Strategic targeting**
   High-quality sites selectively block training bots over search bots
3. **Misinformation asymmetry**
   Conspiracy, propaganda, pseudoscience stay open for AI crawlers
4. **Ideological sorting**
   Centrist outlets block most; both left and right extremes block least

⇒ Evidence from **9,600 news and media sites** with two complementary quality measures: expert ratings and web-graph metrics

Institutional Background

robots.txt

- Voluntary protocol since 1994
- Sites specify which bots to block
- Advisory, not legally enforced
- Can target *specific* AI crawlers

AI Crawlers

- **Training**: GPTBot, Google-Extended, ClaudeBot
  Extract content to build models
- **Search**: ChatGPT-User, PerplexityBot, Applebot-Extended
  Retrieve content to answer queries — can drive traffic back

Key events:

Why Media? Data & Sample

Why media & news?

- News content shapes public discourse and how AI models handle factual claims
- Expert quality ratings exist—enabling direct measurement of adverse selection
- Center of active policy debate (NYT v. OpenAI, EU AI Act)
- **High stakes**: if credible outlets exit but misinformation stays, AI inherits a distorted information environment

Sample

- **9,611** media/news sites (cross-section, February 2026)
- **7,002** sites $\times$ 45 months panel (August 2022–February 2026)
- Source: Media Bias/Fact Check (MBFC)
- robots.txt scraped monthly via HTTP Archive

Quality Measures: Two Complementary Indices

Content quality index *("inside-out")*

Expert-rated: what the site *produces*

- MBFC factual reporting (1–6)
- MBFC credibility rating (1–4)
- Wikipedia notable (binary)

→ PCA first component, standardized

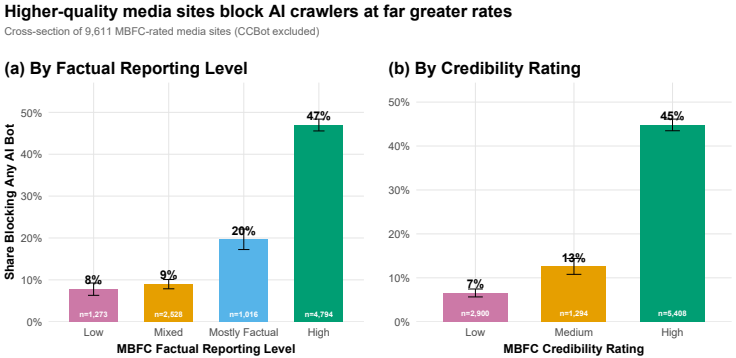
Domain quality index *("outside-in")*

Algorithmic: how the *web* treats the site

- Trust Flow (Majestic, 0–100)
- Open PageRank (web centrality)
- Tranco rank (aggregated popularity)

→ PCA first component, standardized

Result 1: Best Content Most Likely to Leave the Training Pool

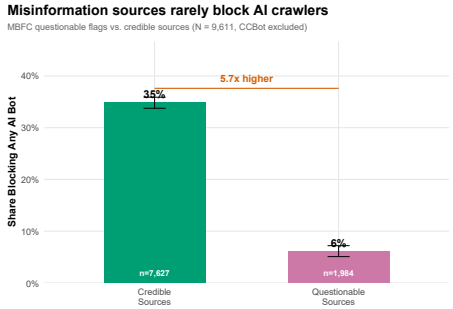


Textbook adverse selection:

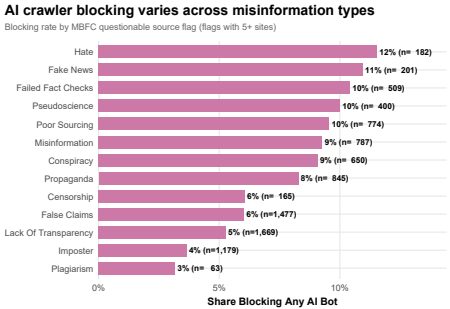
- Low-factual sites: ∼**8%** block; high-factual sites: ∼**47%** block—a **6**× gap
- Monotonically increasing across both factual reporting and credibility
  Consistent across all four quality measures: factual, credibility, content quality index, domain quality index

Result 2: Misinformation Stays Freely Available for AI Training

(a) Credible vs. Questionable



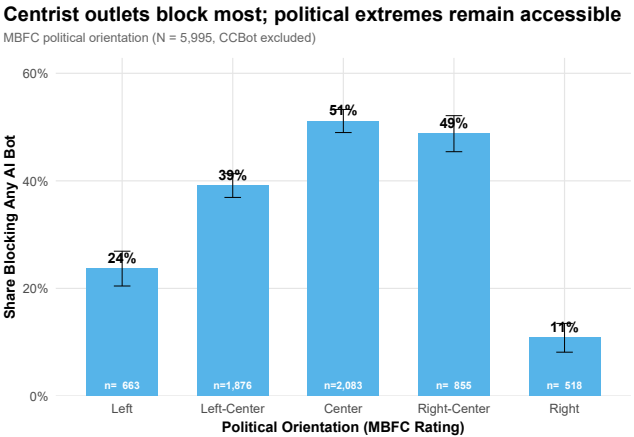
(b) By Misinformation Flag



Opt-out skews the corpus toward misinformation:

- Credible outlets block at **5.7**$\times$ the rate of questionable sources (35% vs. 6%)
- Conspiracy, pseudoscience, propaganda sources: all below 10%
- As credible sources exit, the remaining corpus tilts toward misinformation

Result 3: AI Training Data Is Becoming Ideologically Sorted



Political extremes stay in the training pool:

- Center: **51%** block
- Left: 24% block
- Right: only **11%** block

$\Rightarrow$ Inverted-U: centrist voices exit, extremes remain

Result 3: Ideological Gaps Persist After Quality Controls

	(1) Bias	(2) +Dom. Q	(3) +Cont. Q	(4) Full
Left	−.354***	−.352***	−.234***	−.230***
	(.024)	(.024)	(.026)	(.026)
Left-Ctr	−.173***	−.177***	−.159***	−.163***
	(.017)	(.017)	(.017)	(.017)
Right-Ctr	−.101***	−.102***	−.045**	−.045**
	(.022)	(.022)	(.022)	(.022)
Right	−.483***	−.467***	−.206***	−.186***
	(.028)	(.028)	(.035)	(.035)
Cont. Q			.234***	.236***
			(.018)	(.018)
N	4,650	4,650	4,650	4,650
R^2	.085	.091	.117	.125
Cat. FE				X

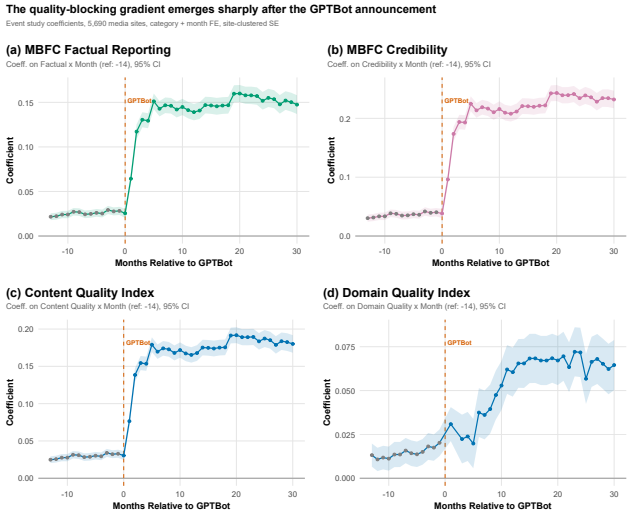
Ref: Center. * $p < .1$, ** $p < .05$, *** $p < .01$

Quality explains part, but not all:

- Right: **−19 pp** vs. Center after full controls
- Left: **−23 pp** vs. Center after full controls
- Content quality is a strong predictor (+24 pp per SD)

⇒ Political orientation has an **independent** effect beyond quality

Result 4: Quality–Blocking Gradient Amplified After GPTBot

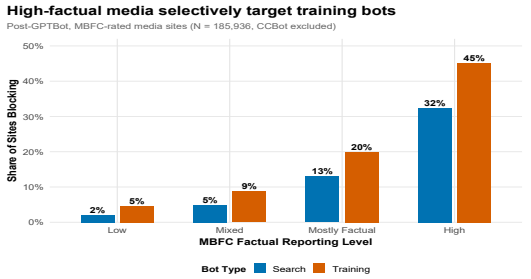


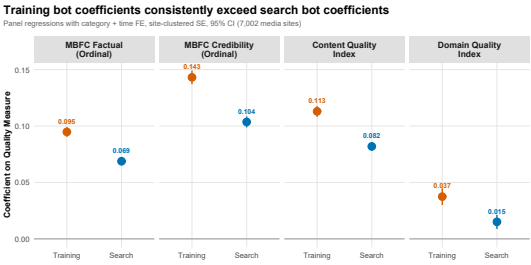
When did adverse selection emerge?

- Pre-period: small, flat quality–blocking gradient
- Aug 2023 (GPTBot): **sharp break**—coefficients jump **5–6**×
- Gradient **keeps widening**: social learning and cascading adoption

⇒ Not a pre-existing pattern—GPTBot made opt-out salient and actionable

Result 5: Adverse Selection Concentrated in Training Data



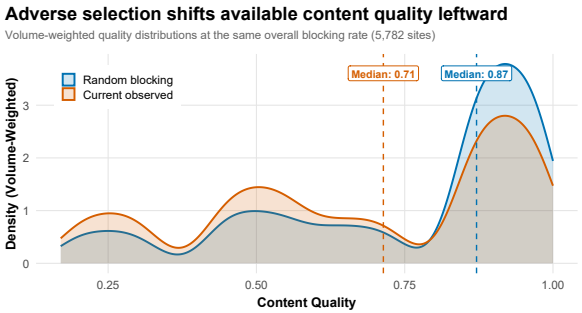


Blocking is strategic, not indiscriminate:

- Quality gap larger for training bots than search bots across factual levels
- Regression coefficients **2–3**$\times$ larger for training bots across all quality measures
- Robust across cross-section, panel, site + time FE

$\Rightarrow$ Producers block value *extraction*; permit value *creation*

Result 6: Counterfactual Simulation — Selection, Not Volume, Drives Quality Loss



Decomposing the quality decline:

- **Counterfactual**: same blocking rate: quality-correlated blocking vs random selection
- Adverse selection accounts for the **entire** quality decline, not blocking per se
- Projections through **2028**: degradation compounds to **18–32%** quality decrease

$\Rightarrow$ Quality loss driven by *selection*, not by volume of opt-outs

Policy Implications

1. **Licensing frameworks**: Replace binary opt-out with a market for data licenses
   Compensate quality producers to stay in the commons
2. **Collective action**: Lower transaction costs via industry-wide standards
   cf. ASCAP for music — centralized licensing reduces bilateral negotiation costs
3. **Regulatory design**: Opt-out regimes must account for adverse selection
   EU AI Act, copyright reform

Opt-out regimes by design produce a lemons equilibrium

Conclusion

Three contributions:

1. **Economics of data markets**: First evidence of adverse selection in AI training data
   No price mechanism → best content exits first → lemons problem (Akerlof, 1970)
2. **AI data governance**: Training data composition is *endogenous*, not given
   The commons degrades by design — producers' strategic opt-out decisions shape what remains
3. **Misinformation & media**: A new channel for AI to amplify misinformation
   Not by generating it, but through the systematic withdrawal of high-quality source material

Conclusion

Three contributions:

1. **Economics of data markets**: First evidence of adverse selection in AI training data
   No price mechanism → best content exits first → lemons problem (Akerlof, 1970)
2. **AI data governance**: Training data composition is *endogenous*, not given
   The commons degrades by design — producers' strategic opt-out decisions shape what remains
3. **Misinformation & media**: A new channel for AI to amplify misinformation
   Not by generating it, but through the systematic withdrawal of high-quality source material

As the best sources withdraw, AI training data becomes systematically **less factual, less credible, and more ideologically skewed**—without institutional solutions, this degradation becomes **self-reinforcing**.

Thank You!

Questions and Comments?

Kai Zhu

Bocconi University