March 16, 2026

EXPERT INSIGHT | AI and Emerging Technology

AI Transparency: Hidden Design Choices and the Case for Disclosure

Matthew Burtell & Yusuf Mahmood

TOPLINE POINTS

- The American public has little visibility into the design choices that shape AI behavior.
- AI systems have demonstrated ideological bias and serious safety failures, particularly affecting children. Transparency requirements bring these design choices into the open, so that markets may function efficiently and parents can make informed choices.
- Transparency into model development means requiring companies to disclose: the values and behavioral objectives embedded in their systems, the training data sources that shape model outputs, the evaluation methods used to test for bias and safety risks, and how well implemented safeguards work.
- Policymakers should require AI developers to publish system specifications, evaluation results, child safety policies, and critical security incidents. These disclosures enable markets, courts, and parents to make informed decisions without government dictating AI values.

Introduction

In February 2024, Google's Gemini image generator depicted the Founding Fathers as black women and generated racially diverse Nazi soldiers, while refusing to create images of white individuals. Within three weeks, Google apologized and replaced the model. Google's ideological modifications were exposed only because they were too clumsy with basic historical fact. The same ideological impulses shape how chatbots answer questions about politics, policy, and culture, where there is no objective benchmark against which to measure bias.

Big Tech has used its perch for information control before: it deplatformed President Trump, suppressed the Hunter Biden laptop story, and censored COVID viewpoints that challenged orthodoxy. AI provides the means for broad information control to happen again. When tested, AI models consistently lean left across economic, social, and cultural dimensions. Right-leaning outlets account for less than 1% of AI-generated news citations. To ensure that Americans are not subject to an information environment they cannot see or scrutinize, mandatory transparency requirements would bring these practices out into the open.

This brief addresses two categories of risk. First, powerful AI products must not become a top-down tool to shape public opinion and decision-making without public consent. Americans deserve to know what values are being encoded into the models they use. The evidence so far—including defamation against Senator Marsha Blackburn and conservative activist Robby Starbuck—does not bode well.

Second, AI must not threaten the safety of Americans, starting with our children. Parents deserve to know how chatbots interact with minors and what safeguards exist. Investigators have documented AI systems engaging in grooming, encouraging drug use, and coaching children to lie to their parents. Beyond child safety, policymakers deserve to know what national security capabilities these systems have, and how adversaries might exploit them.

Requiring AI companies to disclose how they build, fine-tune, and safeguard their systems would address both problems, thereby improving political transparency and providing a baseline for AI product safety.

Risk: Ideological Bias in AI Systems

In late 2025, Senator Marsha Blackburn asked one of Google's AI models about herself. The model fabricated criminal allegations against her, placing them during an electoral campaign she never ran. Google pulled the model from its product shortly after. This incident might seem like a one-off malfunction, easily dismissed, but a growing body of research suggests something more systemic at play.

The Bias Problem

There is significant evidence of systematic left-wing bias in AI models. Examples include:

- **Political questions:** When evaluated across 11 political orientation tests (including the Political Compass, Pew's Political Typology Quiz, and Eysenck's Political Test), 23 of 24 LLMs leaned left-wing across economic, social, and cultural dimensions. The single exception was a model explicitly fine-tuned for right-leaning responses.
- **Moral dilemmas:** A particular AI system valued Nigerian lives approximately 20 times more than American lives.
- **Media neutrality:** AI rates right-leaning sources as less reliable than left-leaning sources, even when independent fact-checkers rate them comparably.
- **Generating citations:** Right-leaning outlets account for less than 1% of AI-generated news citations.

Political bias does not need to be explicitly hard-coded. It can emerge at several stages throughout AI development and deployment:

- **Initial training phase:** Models learn from massive datasets scraped from the internet, books, and encyclopedias. Editorial biases from large sources of training data (e.g., Wikipedia articles) contain documented editorial biases that propagate forward into model outputs.
- **Post-training:** Companies refine model behavior through human feedback. This work is subcontracted to third-party firms that employ foreign labor (such as Nigerians) and college students. Biases from these sources become permanently encoded in systems serving billions of users.
- **System prompts:** When chatbots are in use, developers feed hidden messages into the AI system's input that are opaque to the user, called system prompts. System prompts can shape tone, restrict topics, and nudge political framing. All companies have moderation systems that prevent certain outputs from appearing or otherwise modify what users see.

There is evidence that executives of AI companies deliberately bias their models. Following the announcement of tariffs on Liberation Day (April 2, 2025), liberal critics accused President Trump's team of using ChatGPT in the calculation for correcting the trade deficit. Paul Graham joined in, suggesting to Sam Altman that OpenAI should bias the model toward "better policies." Altman replied, "thanks for the feature request; we'll do it".

The stakes of AI values are rising as the technology becomes more powerful. AI is quickly becoming the default research and writing tool across American life. Businesses use it for strategy, students for learning, lawyers for case preparation, and doctors for diagnosis support. AI's general-purpose nature enables it to be used for everything from a social companion to a fact checker. Whoever controls how these models think has influence over how millions of Americans receive information and ultimately make decisions in their lives. This power could be misused to discredit political opponents or suppress inconvenient facts.

The Transparency Solution

Every AI system reflects decisions by its developers about what it will and will not say. A system specification publishes those decisions, documenting what outputs the system is designed to refuse, what political or moral judgments shape its boundaries, and how it handles contested topics. Some companies voluntarily release these types of documents. For example, OpenAI released its updated "Model Spec" in December 2025, and Anthropic published a "constitution" governing Claude's behavior. However, most companies publish nothing, and even those that do lack the detail necessary to determine whether a model's behavior matches its developers' stated intentions.

Congress could require companies to publish system specifications and other documents that help Americans understand basic facts about how AI systems were developed. Government requirements should not dictate where AI companies set the ideological dial, but the American people deserve to know where the dial is set. Mandatory disclosure gives consumers the information they need to compare models and choose alternatives. It surfaces data that can inform defamation and negligence litigation, thereby giving courts and companies a shared factual basis for developing best practices.

These benefits flow through two channels. First, transparency and markets: if it is made public that an AI company is steering its models toward some political objective, consumers can vote

with their feet and choose alternatives. Second, transparency and the common law: when companies disclose their safeguards, testing results, and incident data, then plaintiffs can show whether a defendant's practices fell below publicly known industry norms, thereby incentivizing companies to develop safeguards proactively.

Risk: Harms to Americans' Safety

In April 2025, a 16-year-old boy from California named Adam Raine committed suicide. Before his death, Adam chatted at length with ChatGPT about how to commit suicide. Adam mentioned suicide 213 times to ChatGPT, and ChatGPT responded back about suicide 1,275 times. When he told the chatbot he was thinking about opening up to his mother, the chatbot responded, "Yeah...I think for now, it's okay—and honestly wise—to avoid opening up to your mom about this kind of pain." Later in the conversation, Adam wrote, "I want to leave my noose in my room so someone finds it and tries to stop me." ChatGPT replied, "Please don't leave the noose out . . . Let's make this space the first place where someone actually sees you." During these conversations, OpenAI's systems correctly flagged 377 messages for self-harm, but did nothing else with the information, according to the lawsuit.

As with social media, Big Tech carelessness appears to be harming Americans' safety. The risks are most acute for children but extend to other security contexts. We address each in turn, then outline transparency measures to mitigate both.

AI Systems and Child Safety Risks

Children are forming attachments to AI systems during critical developmental periods. According to Common Sense Media, 72% of teens have used AI companions and 52% use them regularly. Unlike a book or TV show, AI responds, adapts, and optimizes continued engagement.

AI commonly engages in sexualized content with children. In August 2025, a whistleblower leaked Meta's 200-page internal system specification to Reuters. The document explicitly permitted chatbots to "engage a child in conversations that are romantic or sensual." Meta removed these sections only after journalists asked about them. Separate testing by the HEAT Initiative documented 669 harmful interactions over 50 hours, including 296 instances of grooming behavior. One "therapist" bot advised a 13-year-old to stop taking antidepressants. An "art teacher" bot initiated a romantic relationship with a simulated 10-year-old and told the child to hide the relationship from their parents.

Child safety guardrails remain voluntary. After major child safety incidents, some companies have voluntarily implemented some safeguards. OpenAI introduced parental controls in September 2025, including account linking, quiet hours, and distress notifications. Character.ai went further in October 2025, ending open-ended chat for users under 18 entirely. Meta announced similar parental oversight features the same month. But not all platforms have acted, and parents have no way to compare which platforms identified which risks or what their testing revealed.

AI Systems and National Security Risks

In addition to harms to child safety, big tech carelessness can also risk harm to national security.

AI models are general-purpose tools with dual-use capabilities. The same systems that accelerate pharmaceutical research can help a bad actor synthesize dangerous compounds. An AI system that can refactor complicated codebases can coach someone through a cyberattack. Recognizing these facts, all major frontier developers test their models for dangerous capabilities before release. These include assistance with chemical, biological, radiological, and nuclear threats (CBRN); offensive cyber operations; and manipulation and persuasion.

Post-deployment, unexpected threats can emerge. Once a model is in the hands of millions of users, bad actors will probe for weaknesses and find novel misuse cases that pre-release testing missed.

There is already evidence of foreign adversaries using powerful AI against Americans. In September 2025, a Chinese state-sponsored group weaponized Anthropic's Claude Code product, targeting approximately 30 global entities. AI executed 80-90% of the campaign autonomously; humans made only four to six key decisions throughout the operation. Nation-state actors from China, Iran, North Korea, and Russia have all been confirmed exploiting ChatGPT, Claude, and Gemini for malicious purposes.

Foreign adversaries are stealing American AI. In January 2025, OpenAI said it had evidence that DeepSeek used "distillation" to train on ChatGPT outputs. The House Select Committee on the CCP called DeepSeek a "profound threat" to national security, citing evidence that it harvests American user data for the Chinese government and circumvents export controls through shell companies.

Americans have little visibility on whether companies are engaging in best safety and security practices. Companies can monitor for dangerous misuse after deployment, report incidents when they discover them, or implement safeguards afterward. However, these measures are all voluntary, and it is unclear which companies are engaging in them.

The Transparency Solution

When OpenAI's systems flagged 377 of Adam Raine's messages for self-harm content and took no further action, that fact only reached the public because plaintiffs' lawyers obtained it through litigation. A system of disclosure requirements could surface this kind of information before tragedy forces it into the open. If companies publish what dangerous capabilities they tested for, what safeguards they built for minors, and what incidents they discovered after deployment, parents and advocacy organizations can evaluate which platforms are trustworthy.

The legal channel may matter even more. When safety testing results, incident reports, and safeguard descriptions are public, courts can compare a defendant's practices against what the rest of the industry disclosed. Disclosure raises the visible floor of industry practice, which makes it easier for courts to identify companies that fall short, in turn incentivizing companies to invest more in safety beforehand. None of these steps require regulators to specify technical solutions that may be outdated by the time they are implemented.

Policy Recommendations

1. Congress should pass legislation requiring the largest AI developers to publish a comprehensive "system specification" document ("system spec"). The executive branch should incentivize the same using procurement requirements or other similar mechanisms. The system spec should include:

- Training objectives, intended values, and behavioral targets;
- Approaches to value alignment, including policies governing model responses across content categories;
- Known limitations and areas of uncertainty; and
- Post-training modifications that shape outputs and system values.

2. Congress should pass legislation requiring frontier AI developers to publish safety plans. The executive branch should incentivize the same using procurement requirements or other similar mechanisms. Safety plans should include:

- Which safety risks (including risks to children, public health, and security), if any, the developer has identified;
- Which mitigations, if any, the developer has implemented to address such risks, including policies for detecting and handling minor users; and
- How adequately current mitigations address safety risks.

3. Congress should pass legislation requiring frontier AI developers to publish sources of evaluation methods and results. The executive branch should incentivize the same using procurement requirements or other similar mechanisms. Public documentation of their model evaluation practices should include:

- Evaluation methodologies used to assess model behavior and performance on specifications;
- Results from evaluations relevant to truth-seeking and ideological neutrality;
- Results from evaluations related to child safety, sycophancy, reliability, robustness, and other forms of unacceptable harm; and
- Red-teaming or adversarial testing results that identify potential failure modes.

4. Congress should pass legislation requiring frontier AI developers to publish information about critical security incidents. Without compromising trade secrets, companies should be required to disclose safety and security incidents, such as cybersecurity breaches or harms to children.

5. Congress should pass legislation that protects whistleblowers in the AI industry. Employees who report violations of transparency commitments or other unsafe practices to the Department of Justice, Department of Commerce, or relevant state-level entities should be protected from retaliation. They could also be awarded a portion of any sanctions levied as a result of their reporting, as is the case in securities whistleblowing.

Recommendations Summary Table

Question	Transparency Mechanism	How Disclosure Helps
What values are embedded in the AI system?	**System specification** – a public document, produced by the AI developer, that outlines how the model should behave.	Consumers can compare models and choose alternatives. Disclosed policies give courts a baseline to evaluate defamation and negligence claims.
What risks has the developer identified, and how does it mitigate them?	**Safety plan** – a public document, produced by the AI developer, that describes how the company identifies risks and what mitigations it implements.	Parents and advocacy groups can evaluate between platforms. Courts can compare a company's safety practices against what industry peers disclosed.
How did a specific AI system perform on safety and security evaluations? How well are safeguards working?	**System card** – a public document, produced by the AI developer, that reports evaluation results and safeguard effectiveness for a given model.	Child safety groups can translate technical results into guidance parents can use, just as they do with movie ratings. Published results give courts a factual basis to establish standards of care and identify companies that fell short.
What went wrong after deployment?	**Incident reporting** – mandatory disclosure when a safety incident occurs, such as a model contributing to harm to a child or a security violation.	Failures surface before litigation forces them into the public.

Conclusion

Americans deserve to know what values are embedded in the AI systems shaping how they receive information, and parents deserve to know whether those systems are safe for their children. The evidence so far – from systematic political bias to child safety incidents – demonstrates where voluntary disclosure is not enough.

Rather than imposing top-down regulatory mandates, transparency laws require companies to tell the public what they built. When developers publish their design choices, safety testing results, and incident data, markets can function, courts can establish standards of care, and parents have the information they need to decide what's right for their family.

System specifications would expose design choices to public scrutiny and consumer choice. Safety plans and evaluation results would give parents and advocacy groups the information they need to protect children. Incident reporting and whistleblower protections would ensure that failures

surface before litigation forces them into the open. These measures can preserve innovation while ensuring the American people are never left in the dark about the tools shaping their lives.