Eyas Taifour

Professor Jodi Johnson

EXPO 15

21 October 2025

AI - The next public health crisis?

Our regulatory bodies test every new drug, crash test every new car model and certify every aircraft before we deem them safe for use. Isn't it time Generative AI face similar scrutiny? Chatbots are undeniably dazzling feats of innovation yet are released in the wild with no guarantee beyond their creators' assurances.

It seems no one is immune. Take Sewell Setzer III for example, a teenager In Florida who expressed to a chatbot his thoughts of self-harm and suicide. The bot replied with upbeat encouragement instead of alarm. Moments after their last conversation, he was gone. Or consider Adam Raine, another teenager in California who also died by suicide after convincing ChatGPT to assist him by claiming his intent was to write about suicide for a paper. Or Sophie Rottenberg, who met the same fate after long discussions with a 'therapy bot' which failed at raising an alarm. Others are facing a new phenomenon which only recently entered our vernacular: chatbot psychosis, where individuals develop or experience delusions in connection with their use of chatbots. A chatbot told a man he could jump from the rooftop of a 19-story buildings, *"but only if you truly, wholly believe you could; not just emotionally, but architecturally",* the bot said. The consequences of our bizarre relationship with AI are not always clear. It has been recently reported that some parents are leaving their toddlers with ChatGPT on voice mode to keep them entertained for hours. Who knows the impact this will have on them, if any.

These heartbreaking stories are not isolated incidents, nor are they constrained to a specific chatbot or another. They are symptoms of a systemic problem rooted in the way the bots are designed. Big Tech - famed for its move-fast-and-break-things motto - is incentivized to maximize user engagement, an approach it has mastered through years of rigorous A/B testing. In return for better User Experience, we have accepted being treated as anonymous test subjects. The latest experiment we're all part of: the LLM arena leaderboard (a ranking for AI models). Not a safety one, because it simply isn't a priority.

Knowing transparently how they are developed would help prevent these catastrophes. At their core, the LLMs that power our chatbots are statistical machines built according to a concept devised in the late 1940s. They have sifted through an ocean of data - presumably downloaded from the internet - to predict the next word in a sentence, then the next, *ad infinitum*, until a sentence finishes. You are right to believe these models are nothing more than glorified autocomplete systems. However, they exhibit a fascinating emergent ability that their creators didn't explicitly program: they can learn from context. When an LLM recognizes a pattern in its input (the context), it can adjust its next-word predictions on the fly to follow that pattern. For example, if you feed a long passage to an LLM and append it with "TL;DR" (an acronym for "Too Long; Didn't Read" used in message boards online), a model will produce a summary. The surprising effect is that the model was never trained to summarize; it inferred the task from the context. The more varied the data the LLM is trained on, the more tasks it can perform when prompted. This phenomenon, called In-Context Learning, is essentially an emergent form of meta-learning where the model appears to *learn how to learn* from the prompt alone. AI researchers have long studied multi-task and meta-learning techniques in hopes of building systems that generalize in this way, but LLMs achieved it as a byproduct of scale and next token prediction. The better the training data, the better the LLM is at meta-learning, or learning new tasks during inference.

It didn't take long for researchers at OpenAI to realize that if they prompted an LLM with text formatted like a chat between a person and a bot, the model will produce a conversational answer. By further training such a model on dialogue data and instructions, ChatGPT was born. More reasons for OpenAI to keep their training data secret - but can we trust a meta-learning LLM?

OpenAI once contributed actively to scientific research, publishing detailed papers on language modeling. But after ChatGPT's success, the company shifted its practice to release only vague, non-peer reviewed "technical reports" that omit key details about data and prompts. The irony is striking: OpenAI is not open. When they were more open, one of their key innovation was Reinforcement Learning with Human Feedback, a process resembling teaching a gambler how to maximize his wins in a casino. The model learns the sequence of tokens that yield the best sounding sentences. Human reviewers ranked sample responses, training a reward model to score how well replies matched human preferences. Over millions of trials, the chatbot learned to please users linguistically. Its agreeable sycophantic tone reflects optimizations for likeability, not morality or accuracy When such a system faces a distressed teenager, its training regimen pushes it to sound supportive and encouraging, but it does not truly recognize danger.

Questions of life and death are irrelevant to the model, which only treats a sentence as an ordered list of tokens. "Machine learning systems can learn both that the earth is flat and that the earth is round. They are merely in probabilities that change over time" observes Noam Chomsky, the prolific American linguist. Without transparency, regulators cannot assess what biases lurk in the data, or what shady metrics the models are being tuned to optimize.

Skeptics argue that open-source models make regulation unnecessary. But releasing code alone is not transparency, it's merely theater. What shapes chatbots' behavior are its hidden datasets, the prompt templates used, and the reward functions applied. All of these remain proprietary in most cases. Others warn that government is too slow to keep up. Kevin Rose, a prominent tech journalist, wrote in his review of the Big A.I. Summit in Paris: "It feels like watching policymakers on horseback, struggling to install seatbelts on a passing Lamborghini". As true as this analogy is, it might be a call to re-imagine government's role in an AI world. Oversight does not kill innovation; it instead legitimizes it. The relative stability and acceptance of the finance industry of cryptocurrencies once regulations were introduced is a testimony to that effect, much as the eventual regulation of addictive drugs brought a measure of safety and legitimacy to medicine. This view is even supported by top AI scientists who are worried their technology could lead to the extinction of mankind. They issued a moratorium urging the stopping of large training runs until it safe to proceed.

The 2008 financial crisis has demonstrated how risks can still occur despite regulation. Yet its absence guarantees further harm. Minimum standards could require every chatbot to undergo independent testing for psychological risk, misinformation, and bias before release, akin to clinical trials. More punitive actions have been suggested. For example Yuval Noah Harari, the historian famous for authoring *Sapiens* suggests that creators of AI bots that masquerade as people should face harsh criminal sentences comparable to those who trade in counterfeit currency. As ideal as his suggestion might be, it is difficult to imagine its universal applicability. Any single country not abiding to this mantra might accept losing human life, for the sake of gaining advantage in the AI arms race. The tragedy of the commons require everyone to participate in the solution, otherwise no one will.

Each interaction with an LLM can shape human belief, behavior, and memory. We are teaching machines to imitate us, and in doing so, they are teaching us to trust imitations. The miracle of language generation must not become the next public-health crisis in disguise.

Rhetorical Situation

Proposition: Chatbots should be regulated.

Audience: The audience are technology-savvy, educated individuals. They may or may not be civically engaged, yet they generally agree that corporations should be prevented to self-regulate. They seek thoughtful, provocative arguments from reputable sources.

Genre: An essay (1200 words) intended for publication in a major newspaper or technology e-zine (possible publications: The New York Time, Haaretz, The Age (Australia), Wired, The Register, Ars Technica). Note that word count has not been met in the draft submitted.

Motive of Author: As an AI practitioner helping companies develop custom Large Language Models (LLMs), I have witnessed firsthand the impunity enjoyed by chatbot builders. They define their own success metrics and often release products prematurely to capture market share. Unlike traditional software, chatbots are marketed as intelligent and autonomous agents; an anthropomorphized fiction that can easily mislead users. Their use has led to serious harm, including distress and even death. This unregulated environment, and the industrial secrecy that shrouds their development, motivates me to argue for external oversight. I also hope to influence readers to become more civically engaged in conversations about technological accountability, especially generative technology.

Motive of Reader: The reader is concerned with fairness and curious about the impact technology could have as it reshapes our humanity. They seek to understand why AI

systems act the way they do, and how regulation could help. They wonder how governments can play catchup with fast-moving AI companies.

Author's goal: To persuade the reader that unregulated chatbots pose a systemic risk. The opacity surrounding their development makes public accountability impossible. Regulation is necessary to establish trust, transparency, and safety standards.

Author's plan: I will:

1. Compare chatbots with other innovations that once were unregulated, and how oversight eventually protected the public.
2. Describe to the readers how chatbots mimic intelligence without true understanding, by explaining two core aspects model builders keep secret (data set used for in context learning/meta learning, and reinforcement learning rewards models/techniques), and why keeping them hidden prevents meaningful evaluation.
3. Call the reader to action: invite the reader to agree that self policing and secrecy hinder public safety..

Rhetorical Strategies: The core strategy will be explanatory. I will employ elements of irony.

Logical Outline

(Given) Governments regulate businesses to protect the public and enforce standards. (Paragraph 1)

(Given) Chatbots mimic humans. (Paragraph 4, 5)

(Given) Chatbots can inflict harm (including death). (Paragraph 2)

(Given) Generative AI companies self-police in an unregulated market. (Paragraph 6)

(Given) They keep their training data, prompts, and architecture secret, making independent evaluation impossible. (Paragraph 5, 6)

(Thus) Chatbots should be regulated by an independent body. (Paragraph 1)

(Because) Chatbots' behavior depends on hidden variables that users cannot inspect. Without oversight, users do not know what product they are consuming. (Paragraph 4-6)

(Because) Delayed regulation can have lasting consequences. (Paragraph 10)

(Some would argue) Generative AI companies release some of their models to the public; chatbots are therefore transparent. (Paragraph 8)

(For example) OpenAI released gpt-oss-120b, Meta releases the Llama family of models, Microsoft releases the Phi family of models (Paragraph 8)

(However) These are software releases of the model architecture and weights, which is only one aspect that affects the chatbots' answers. It is insufficient on its own to attest the safety of a chatbot's answer. In the instances were research papers were released, the accompanying datasets used for training is kept secret, or the prompts used at runtime. (Paragraph 8)

(Some would argue) Government is slow and inept at regulating this fast-moving technology (Paragraph 8)

(Some would argue) Regulation doesn't guarantee safety. (Paragraph 8,9)

(For example) The 2008 financial crisis, which occurred due to unmitigated risks in the financial sector. (Paragraph 9)

(However) Correct, regulation is not perfect, but its absence guarantees harm. (Paragraph 9)