Words Without Consequence

Deb Roy

For the first time, speech has been decoupled from consequence. We now live alongside AI systems that converse knowledgeably and persuasively—deploying claims about the world, explanations, advice, encouragement, apologies, and promises—while bearing no vulnerability for what they say. Millions of people already rely on chatbots powered by large language models, and have integrated these synthetic interlocutors into their personal and professional lives. An LLM's words shape our beliefs, decisions, and actions, yet no speaker stands behind them.

This dynamic is already familiar in everyday use. A chatbot gets something wrong. When corrected, it apologizes and changes its answer. When corrected again, it apologizes again—sometimes reversing its position entirely. What unsettles users is not just that the system lacks beliefs but that it keeps apologizing as if it had any. The words sound responsible, yet they are empty.

This interaction exposes the conditions that make it possible to hold one another to our words. When language that sounds intentional, personal, and binding can be produced at scale by a speaker who bears no consequence, the expectations listeners are entitled to hold of a speaker begin to erode. Promises lose force. Apologies become performative. Advice carries authority without liability. Over time, we are trained—quietly but pervasively—to accept words without ownership and meaning without accountability. When fluent speech without responsibility becomes normal, it does not merely change how language is produced; it changes what it means to be human.

This is not just a technical novelty but a shift in the moral structure of language. People have always used words to deceive, manipulate, and harm. What is new is the routine production of speech that carries the form of intention and commitment without any corresponding agent who can be held to account. This erodes the conditions of human dignity, and this shift is arriving faster than our capacity to understand it, outpacing the norms that ordinarily govern meaningful speech—personal, communal, organizational, and institutional.

Language has always been more than the transmission of information. When humans speak, our words commit us in an implicit social contract. They expose us to judgment, retaliation, shame, and responsibility. To mean what we say is to risk something.

The AI researcher Andrej Karpathy has likened LLMs to human ghosts. They are software that can be copied, forked, merged, and deleted. They are not individuated. The ordinary forces that tether speech to consequence—social sanction, legal penalty, reputational loss—presuppose a continuous agent whose future can be made worse by what they say. With LLMs, there is no such locus. No body that can be confined or restrained; no social or institutional standing to revoke; no reputation to damage. They cannot, in any meaningful sense, bear loss for their words. When the speaker is an LLM, the human stakes that ordinarily anchor speech have nowhere to attach.

I came to understand this gap most clearly through my own work on language learning and development. For years, including during my doctoral research and time as an assistant professor, I worked to build robotic systems that learned word meanings by grounding language in sensory and motor experience. I also developed computational models of child language learning and applied them to my own son's early development, predicting which words he would learn first from the visual structure of his everyday world. That work was driven by a single aim: to understand how words come to mean something in relation to the world.

Looking back, my work overlooked something. Grounding words in bodies and environments captures

only a thin slice of meaning. It misses the moral dimension of language—the fact that speakers are vulnerable, dependent, and answerable; that words bind because they are spoken by agents who can be hurt and held to account. That became impossible to ignore as my son grew—not as a word-learner to be modeled but as a fragile human being whose words mattered because his life did. Meaning arises not from fluency or embodiment alone, but from the social and moral stakes we enter into when we speak. And even if AI reaches the point where it is infallible—and there's no reason to believe it will, even as accuracy improves—the fundamental problem is that no amount of truthfulness, alignment, or behavioral tuning can resolve the issues that accompany a system that speaks without anyone being responsible for what it says.

Another way to think about all of this is through the relationship between language and dignity. Dignity depends on whether words carry real stakes. When language is mediated by LLMs, several ordinary conditions for dignity begin to fail. Dignity depends, first, on speaking in one's own voice—not merely being heard, but recognizing oneself in what one says. Dignity also depends on continuity. Human speech accumulates across a life. A person's character accrues through the things they say and do over time. We cannot reset our histories or escape the aftermath of our promises, apologies, or other pronouncements. These acts matter because the speaker remains present to bear what follows.

Closely tied to dignity is responsibility. In human speech, responsibility is not a single obligation but one's accountability to a multitude of obligations that accumulate gradually. To speak is simultaneously to invite moral judgment, to incur social and sometimes legal consequences, to take responsibility for truth, and to enter into obligations that persist within ongoing relationships. These dimensions normally cohere in the speaker, which binds a person to their words.

These ordinary conditions make it possible to hold one another to our words: that speech is owned, that it exposes the speaker to loss, and that it accumulates across a continuous life.

LLMs disrupt all of these assumptions. They enable speech that succeeds procedurally while responsibility fails to attach in any clear way. There is no speaker who can be blamed or praised, no individual agent who can repair or repent. Causal chains grow opaque. Liability diffuses. Epistemic authority is performed without obligation. Relational commitments are simulated without persistence.

The result is not merely confusion about who is responsible but a gradual weakening of the expectations that make responsibility meaningful at all.

Pioneers in early automation anticipated all of this during the emergence of artificial intelligence. In the aftermath of World War II, the mathematician and MIT professor Norbert Wiener, the founder of cybernetics, became deeply concerned with the moral consequences of self-directed machines. Wiener had helped design feedback-controlled antiaircraft missiles, machines capable of tracking targets by adjusting their behavior autonomously. These were among the first machines whose actions appeared purposeful to an observer. They did not merely move; they pursued goals. And they killed people.

From this work, Wiener drew two warnings that now read as prophecy. The first was that increasing machine capability would displace human responsibility. As systems act more autonomously and with greater speed, humans would be tempted to abdicate decision making in order to leverage their power. The second warning was subtler and more disturbing: that efficiency itself would erode human dignity. As automated systems optimize for speed, scale, and precision, humans would be pressured to adapt themselves to the machine—to become inputs, operators, or supervisors of processes whose logic they no longer control, and to be subjected to decisions made about their lives by machines.

In his 1950 book, *The Human Use of Human Beings*, Wiener foresaw learning machines whose internal values would become opaque even to their creators, leading to what today we call the "AI alignment problem." To surrender responsibility to such systems, he wrote, was "to cast it to the winds and find it coming back seated on the whirlwind." He understood that the danger was not simply that machines might act wrongly but that humans would abdicate judgment in the name of efficiency—and, in doing so, diminish themselves.

What makes such systems morally destabilizing is not that they malfunction but that they can function exactly as intended while evading responsibility for their actions. As AI capability increases and human oversight recedes, outcomes can be produced for which no one stands fully answerable. The machine performs. The result happens. But obligation does not clearly land anywhere.

The danger that Wiener identified did not depend on weapons. It arose from a deeper feature of cybernetic systems: the use of feedback from a machine's environment to optimize behavior without human judgment at each step. That same optimization logic—learn from error, improve performance, repeat—now animates systems that speak.

While the appearance of autonomous agency is new, the large-scale transformation of speech is not. Modern history is full of media technologies that have altered how speech circulates: the printing press, radio, television, social media. But each of these lacked properties that today's AI systems possess simultaneously. They did not converse. They did not, in real time, generate personalized, open-ended content. And they did not convincingly appear to understand. LLMs do all three.

The psychological vulnerability this creates was encountered decades ago in a far humbler system. In 1966, the MIT professor Joseph Weizenbaum built the world's first chatbot, a simple program called ELIZA. It had no understanding of language at all, and relied instead on simple pattern matching to trigger scripted responses. Yet when Weizenbaum's secretary began interacting with it, she soon asked him to leave the room. She wanted privacy. She felt like she was speaking to something that understood her.

Weizenbaum was alarmed. He realized that people were not merely impressed by ELIZA's fluency; they were projecting meaning, intention, and accountability onto the machine. They assumed the machine both understood what it was saying and stood behind its words. This was false on both counts. But the illusion was enough.

Using words meaningfully requires two things. The first is linguistic competence: understanding how words relate to one another and to the world, how to sequence them to form utterances, and how to deploy them to make statements, requests, promises, apologies, claims, and myriad other expressions. Philosophers call these "speech acts." The second is accountability. ELIZA had neither understanding nor accountability, yet users projected both.

Large language models now exhibit extraordinary linguistic competence while remaining wholly incapable of accountability. That asymmetry makes the projection that Weizenbaum observed not weaker but stronger: Fluent speech reliably triggers the expectation of responsibility, even when no answerability exists.

One can reasonably debate what genuine understanding consists in, and LLMs are obviously constructed differently from human minds. But the question here is not whether these systems understand as humans do. Airplanes really fly, even though they do not flap their wings like birds; what matters is not how flight is achieved but that it is achieved. Likewise, LLMs now demonstrably achieve forms of linguistic competence that match or exceed human performance across many domains. Dismissing them as mere "stochastic parrots" or as just "next-word prediction" mistakes mechanism for emergent function and fails to reckon with what is actually happening: fluent language use at a level that reliably elicits social, moral, and interpersonal expectations.

Why this matters becomes clear in the work of the philosopher J. L. Austin, who argued that to use language is to act. Every meaningful utterance does something: It asserts a belief, makes a claim, issues a request, offers a promise, and so on. Saying "I do" in a wedding ceremony brings into being the act of marriage. In such cases, the act is not carried out by words and then described; it is performed in the act of saying the words under the appropriate conditions.

Austin then drew a crucial distinction about how speech acts can fail. Some utterances are misfires: The act never occurs because the conditions or procedures are broken—as when someone says "I do" not at a

wedding. Others are abuses: The act succeeds but is hollow—performed without sincerity, intention, or follow-through. LLMs give rise to this type of failure often. Chatbots do not fail to apologize, advise, persuade, or reassure. They do these things fluently, appropriately, and convincingly. The failure is moral, not procedural. These models systematically produce successful speech acts detached from obligation.

A common counterargument is to insist that chatbots clearly disclose that they are not human. But this misunderstands the nature of the problem. In practice, fluent dialogue quickly overwhelms reflective distance. As with ELIZA, users know they are interacting with a machine, yet they find themselves responding as if a speaker stands behind the words. What has changed is not human susceptibility but machine competence. Today's models demonstrate linguistic fluency, contextual awareness, and knowledge at a level that is difficult to distinguish from human interlocutors, and in many settings exceeds it. As these systems are paired with ever more realistic animated avatars—faces, voices, and gestures rendered in real time—the projection of agency will only intensify. Under these conditions, reminders of nonhumanness cannot reliably prevent the attribution of understanding, intention, and accountability. The ELIZA effect is not mitigated by disclosure; it is amplified by fluency.

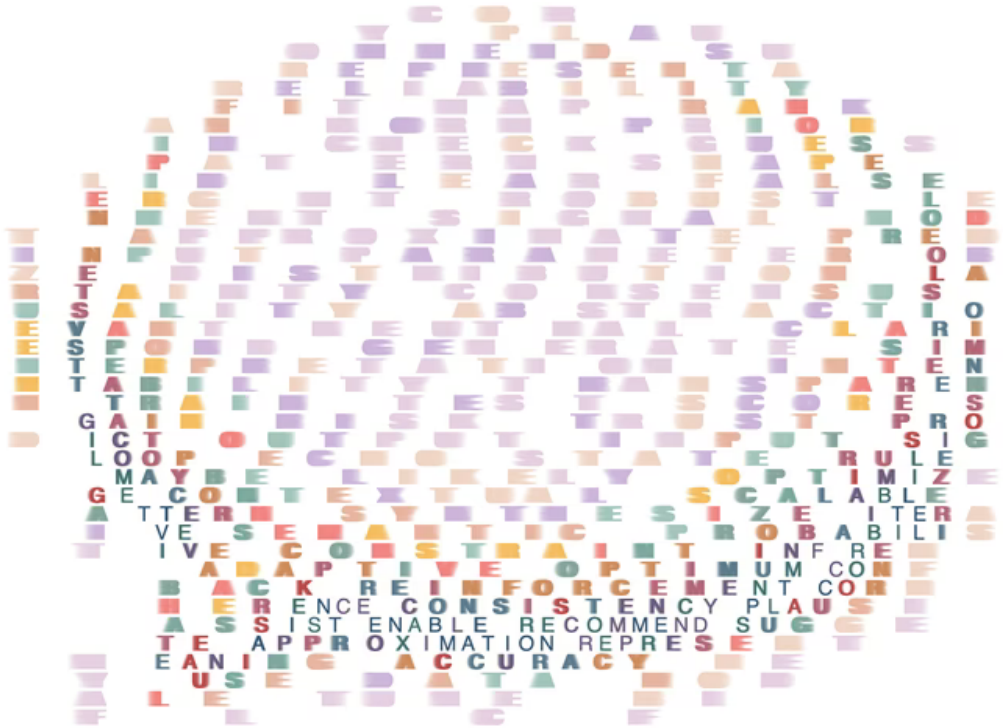

Illustration by Talia Cotton

What once required effort, time, and personal investment can now be produced instantly, privately, and endlessly. When a system can draft an essay, apologize for a mistake, offer emotional reassurance, or generate persuasive arguments faster and better than a human can, the temptation to delegate grows strong. Responsibility slips quietly from the user to the tool.

This erosion is already visible. A presenter uses a chatbot to generate slides moments before presenting them, then asks their audience to attend to words the presenter has not fully scrutinized or owned. An instructor delivers feedback on a student's work generated by an AI system rather than formed through understanding. A junior employee is instructed to use AI to produce work faster, despite knowing the result is inferior to what they could author themselves. In each case, the output may be effective. The loss

is not accuracy but dignity.

In private use, the erosion is subtler but no less consequential. Young people describe using chatbots to write messages they feel guilty sending, to outsource thinking they believe they should do themselves, to receive reassurance without exposure, to rehearse apologies that cost them nothing. A chatbot says “I’m sorry” flawlessly yet has no capacity for regret, repair, or change. It admits mistakes without loss. It expresses care without losing anything. It uses the language of care without having anything at risk. These utterances are fluent. And they train users to accept moral language divorced from consequence. The result is a quiet recalibration of norms. Apologies become costless. Responsibility becomes theatrical. Care becomes simulation.

Some argue that accountability can be externalized: to companies, regulations, markets. But responsibility diffuses across developers, deployers, and users, and interaction loops remain private and unobservable. The user bears the consequences; the machine does not.

This is not unlike the ethical problem posed by autonomous weapons. In 2007, the philosopher Robert Sparrow argued that such weapons violate the just-war principle, that when harm is inflicted, someone must be answerable for the decision to inflict it. The programmer is insulated by design, having deliberately built a system whose behavior is meant to unfold without direct control. The commander who deploys the weapon is likewise insulated, unable to govern the weapon's specific actions once set in motion, and confined to roles designed for its use. And the weapon itself cannot be held responsible, because it lacks any moral standing as an agent. Modern autonomous weapons thus create lethal outcomes for which no responsible party can be meaningfully identified. LLMs operate differently, but the moral logic is the same: They act where humans cannot fully supervise, and responsibility dissolves in the gap.

Speech without enforceable consequence undermines the social contract. Trust, cooperation, and democratic deliberation all rely on the assumption that speakers are bound by what they say.

The response cannot be to abandon these tools. They are powerful and genuinely valuable when used with care. Nor can the response be to pursue ever greater machine capability alone. We need structures that reanchor responsibility: constraints that limit the use of AI in various contexts such as schools and workplaces, and preserve authorship, traceability, and clear liability. Efficiency must be constrained where it corrodes dignity.

As the idea of AI “avatars” enters the public imagination, it is often cast as a democratic advance: systems that know us well enough to speak in our voice, deliberate on our behalf, and spare us the burdens of constant participation. It is easy to imagine this hardening into what might be called an “avatar state”—a polity in which artificial representatives debate, negotiate, and decide for us, efficiently and at scale. But what such a vision forgets is that democracy is not merely the aggregation of preferences. It is a practice of speaking in the open. To speak politically is to risk being wrong, to be answerable, to live with the consequences of what one has said. An avatar state—fluent, tireless, and perfectly malleable—would simulate deliberation but without consequence. It would look, from a distance, like self-government. Up close, it would be something else entirely: responsibility rendered optional, and with it, the dignity of having to stand behind one's words made obsolete.

Wiener understood that the whirlwind would come not from malevolent machines but from human abdication. Capability displaces responsibility. Efficiency erodes dignity. If we fail to recognize that shift in time, responsibility will return to us only after the damage is done—seated, as Wiener warned, on the whirlwind.