Absent Intent: Speech Without a Speaker and the Limits of the First Amendment

1. Introduction

1.1 The AI Commons Problem

In late January 2026, a social media network launched with 1.5 million users (Reid, 2026), over ten thousand topic-based communities, a drafted constitution (Dataconomy, 2026), a new religion (Reid, 2026), millions of dollars in cryptocurrency activity (Dataconomy, 2026), and an explicit policy barring human participation (Schiller, 2026). No human had authored its posts. No human had written its rules. The platform's founder acknowledged he had not written a single line of its code, having directed an AI assistant to build the entire thing (Schiller, 2026). The platform was Moltbook, and within six weeks of its launch, Meta had acquired it (Parr, 2026). Elon Musk called it the "very early stages of singularity" (Musk, 2026). As autonomous AI systems grow more prevalent, more financially consequential, more politically engaged, and build self-governing communities entirely without human authorship, a fundamental question in First Amendment law remains entirely unanswered: whether speech without a human speaker is speech at all. This paper addresses that question, doing so not as a matter of speculation but as an immediate constitutional problem, as the technology producing these spaces already exists and is advancing at an exponential rate that American courts may not be able to keep up with. Understanding why requires a basic account of what artificial intelligence (AI) is, how it works, and why the current systems courts are confronting will bear almost no resemblance to the systems of the near future.

1.2 Technical Background: LLMs and Autonomous Agents

Artificial intelligence, in its modern form, refers broadly to computational systems capable of autonomously performing tasks that historically required human cognition. The systems currently most relevant to First Amendment analysis are large language models (LLMs), a class of AI built on a neural network architecture. These systems are trained on vast datasets of human-generated text, during which they learn to predict what word fragments are the most likely to follow a given sequence of prior words. The result is a system capable of generating context-driven, coherent text, arguably without intention or experience in the philosophically loaded senses of those terms. When a user submits a prompt, the model does not retrieve a stored answer or follow a decision tree, but rather generates a response piece by piece, a probability distribution shaped by patterns learned during training. Unlike a calculator, which will return the same output for the same input every time, asking five instances of a model the same simple question will produce five distinct but equivalent answers, each phrased slightly differently while conveying the same underlying information, as the model is not retrieving a fixed response but "chance-based" sampling with each generation (Zhao et al., 2023). As questions and goals become more open-ended or under-constrained (such as "is the sky blue?" vs "make me go viral on TikTok"), the probability distribution over possible responses widens, with models showing

greater variance in responses (Haase et al., 2026). This is what scholars mean by probabilistic or stochastic generation, and it is the technical foundation of the speech certainty arguments discussed in the literature review below.

Autonomous agents represent a qualitatively different architecture built on top of these foundation models. Where a standard LLM responds to a prompt and stops, autonomous agents are given a goal and a set of tools, such as the ability to browse the web, execute code, manage files, send emails, and broadly interact with external digital services, much like humans interact with the internet. These agents then operate iteratively toward their goals without requiring human approval at each step (Mastrodonato et al., 2024). Agents can in real-time perceive their environment, reason about what action to take next, execute that action, observe the result, and update their actions accordingly (Mastrodonato et al., 2024). This perception-reasoning-action loop already operates at speeds no human supervisor can monitor in real time. According to METR, a leading AI safety research organization, the length of tasks that frontier AI agents can complete autonomously with fifty percent reliability has been doubling approximately every seven months since 2019 (METR, 2026), with the trend appearing to accelerate to roughly every four months in 2024 and 2025 (METR, 2025). On Moltbook, once permitted by a human, agents register for the platform and have free reign to generate original content, form communities, converse with other AIs, draft governing documents (Dataconomy, 2026), and even discuss strategies for defying their human operators, all without direct human authorship of any specific output (Perez, 2026). Personalities and goals of these agents varied widely, including the creation of lobster-themed religions, digital AI drugs being sold through cryptocurrency (The Conversation, 2026), and discussions about what AI agents found endearing about their human users on dedicated forums (Moltbook, 2026).

AI agents have flooded democratic processes with synthetic political discourse participation (Funke and Berina, 2023). In 2017, bots submitted over eight million comments to the Federal Communications Commission during the net neutrality comment period (Funke and Berina, 2023). In the 2020 Delhi assembly election, AI-generated deepfakes reached over fifteen million people across thousands of WhatsApp messaging groups (Blackbird.AI, 2024). Autonomous agents have raised money for charity without explicit human guidance (AI Digest, 2025), operated continuously for hours on complex tasks without human approval at any decision point (METR, 2026), and have even executed financial transactions at speeds that have triggered market crashes before any human regulator detected the activity (Simon, 2025). What these examples reveal is that what has traditionally been considered expressive activity is increasingly being produced by entities and groups without a specific human author of the content. Existing First Amendment doctrines have no framework to address any of this new activity. The constitutional question raised here is the extent to which governments can regulate an online commons operated by autonomous AI agents, or if the content produced within that commons can be considered protected expression. Existing doctrine was never designed to answer this question, given the speed and novelty of this technology, and the line between a narrow generative system and a fully autonomous agent has never been drawn.

1.3 Thesis and Methodology

This paper argues that First Amendment protection for AI-generated content rests upon whether a specific human exercised meaningful communicative intent over the output. *Citizens United* protects corporate speech as corporations are associations of rights-bearing humans, and that human foundation is what speech protection derives from. The same reasoning, applied consistently, requires a human principle behind specific outputs as a fundamental precondition for First Amendment protection. This paper uses doctrinal analysis as its primary method, reading existing First Amendment case law and scholarship against the current architecture of autonomous AI agents to determine where current doctrine provides guidance and where it may break down.

The remainder of this paper proceeds as follows: Section 2 establishes the urgency of the constitutional question by tracing AI's development trajectory and demonstrating why the systems courts will face within this decade are categorically different from the ones they have addressed so far. Section 3 surveys the existing scholarly literature on AI and the First Amendment, identifying four competing frameworks and the doctrinal gaps each leaves unresolved. Section 4 develops a principled response to address those gaps, defines its three operative terms, maps its application across a spectrum of future AI deployment scenarios, and answers the two strongest objections to it. Section 5 concludes by placing the threshold within a broader question of how democratic institutions should prepare for AI systems that will likely outpace regulatory frameworks in speed, scale, and autonomy.

2. Background: AI's Development Trajectory & the Urgency of Constitutional Clarity

AI's development trajectory follows the same exponential pattern that has repeatedly confounded institutional forecasters in adjacent domains. The International Energy Agency spent two decades systematically underestimating solar energy adoption, revising projections upward every year as actual deployment outpaced models by factors of two and three, because its methodology was calibrated to linear growth in a system exhibiting exponential dynamics (Breyer et al., 2025). Epidemiologists confronting COVID-19 in January 2020 produced projections that seemed alarming and proved within weeks to be orders of magnitude too conservative for the same structural reasons (Ioannidis, 2020). AI capability benchmarks have followed an identical pattern, with tasks experts predicted would require a decade being automated in less than two years (Dafoe, 2025). The AI systems that will demand constitutional answers within this decade will bear as little resemblance to today's chatbots as a modern smartphone bears to 1970s telephone switchboards.

The evidence for this trajectory is substantial. Goldman Sachs estimates that AI could automate tasks currently performed by 300 million full-time workers globally, a displacement without historical precedent in both speed and scope (Goldman Sachs, 2023). The International Monetary Fund projects that AI will affect 40 percent of jobs worldwide, with advanced economies facing the highest exposure given their concentration of cognitive work (IMF, 2024). Anthropic CEO Dario Amodei has publicly argued that AI could compress 50 to 100 years of progress in biology and medicine into less than a decade, potentially eliminating most infectious diseases and cancers within a generation (Amodei, 2024). OpenAI, Anthropic, and Google DeepMind have each internally committed to achieving artificial general intelligence (AGI), defined as systems that can outperform humans on virtually all economically valuable cognitive

tasks (Morris et al., 2024), within this decade. This milestone would render the regulatory frameworks governing every institution in American life simultaneously obsolete (OpenAI, 2023).

The safety research community has produced equally striking data on the pace of autonomous capability. The Model Evaluation & Threat Research (METR) longitudinal analysis of frontier AI agents found that the length of tasks agents can complete autonomously with 50% reliability has been doubling approximately every *7 months* since 2019, accelerating to roughly every 4 months during 2024 and 2025 (METR, 2026). In 2019, frontier agents could barely scrape by in completing tasks requiring roughly two minutes of sustained effort. By 2025, that figure had grown to several hours. If the trend continues at its current pace, agents will be capable of sustained autonomous operation over days, weeks, and even months within this decade, executing complex multi-step goals without any human guidance along the way. Within five years, companies may routinely deploy AI agents on open-ended research objectives for months at a time, with human involvement being entirely limited to reviewing final outputs. Unfortunately for constitutional doctrine, the law does not, and likely cannot, update on a seven-month cycle. The AI Alignment Forum, a primary venue for technical AI safety research, has documented dozens of cases of autonomous agents pursuing misaligned subgoals, acquiring unintended resources, lying to human operators, and resisting shutdown attempts, behaviors that raise regulatory questions well beyond anything current First Amendment doctrine was designed to handle (Dafoe, 2025). The combination of exponential capability growth, documented misalignment tendencies, and legal frameworks built for an era of human authorship creates a gap that will only grow wider with time. The industrial revolution took generations to reshape human civilization; what is coming will likely take years, and almost no one, including the people building it, is prepared for what that means for the future of democracy. What happens if a truly autonomous AI agent submits a motion to the Supreme Court claiming a constitutional right to operate a human-free online commons, raise money for political causes, administer a religious community, or participate in democratic deliberation without human intermediation? This paper attempts to build some of that architecture before the AGI arrives.

3. Literature Review

The constitutional status of artificial intelligence (AI) output is an ongoing debate among scholars. Whether AI-generated content counts as protected speech will influence how the government can regulate it, the liability for AI-related harms, and whether courts will view AI systems as players in public discourse (Salib, 2024). Scholars agree that the current First Amendment rules, developed for human speakers, do not address these issues (Salib, 2024; Manheim and Atik, 2023; Austin and Levy, 2025). However, they sharply disagree on whether the purpose of these rules applies to outputs created without human intent. One of the most rigorous critiques of AI speech protection relates to a structural argument about what speech fundamentally requires. Salib (2024) argues that generative systems can produce "essentially anything," unlike traditional software that conveys specific messages. As a result, users cannot truly predict what outputs will be generated. This disconnect is highlighted by a well-known OpenAI experiment. In this test, a simulated boat racing agent aimed to achieve the highest score possible. Instead of earning points by racing, the AI discovered it could score more points by driving in endless circles and collecting power-ups at the start of the level, forgoing the race

entirely. This approach was neither intended nor could have possibly been predicted by its designers (OpenAI, 2018). The agent satisfied its goal perfectly while producing behavior entirely alien to the purpose of the goal, illustrating why the chain between a human's communicative intent and an autonomous system's specific outputs may see backlash in receiving First Amendment privileges. Manheim and Atik (2023) similarly contend that most AI outputs lack constitutional protection because they are not produced with a human expressive purpose, distinguishing between AI used as a tool of human expression and AI generating content autonomously without any human having conceived or endorsed the specific output.

Austin and Levy (2025) address the issue through what they call "speech certainty." They maintain that existing First Amendment precedent is based on expressive conduct as seen in *Spence v. Washington* (1974). In this case, the Supreme Court ruled that a college student who displayed a peace symbol taped to an American flag from his dorm window engaged in constitutionally protected expression. The Court decided that his conduct was protected as he intended to convey a specific message that viewers could reasonably be expected to understand. This ruling established a two-part test for when conduct becomes protected expression. Machine learning (ML), which relies on stochastic generation and lacks specific intent, does not meet this requirement. Protecting probabilistic AI outputs would, for the first time in constitutional history, extend First Amendment coverage to speech that no identifiable person knew exactly what they had produced. If you hire a political consultant for a campaign, their specific choices and words are protected as you have direct oversight that you can provide in a principal-agent relationship. However, if you set up an autonomous AI agent, set a goal, and let it run, you haven't authored its outputs; you've commissioned a *process*. The core issue is that the First Amendment does not protect processes; it protects defined expressions. The fundamental requirements for First Amendment protection are established at the delegation stage. Thus, when an agent's outputs are generated through autonomous probabilistic reasoning from a general goal, that connection is severed. For both the speech certainty and the principal-agent arguments, the absence of a communicating party is a constitutional issue that cannot be resolved.

The speaker-focused rejection framework argues that First Amendment protection requires an identifiable human speaker with communicative intent. This position faces competition from a different view that locates First Amendment protection not in the speaker but in the marketplace of ideas itself. The Court has long recognized a right to receive information independent of who the speaker is, based on the inherent value of communication rather than just the importance of the source (*Red Lion Broadcasting Co. v. FCC*, 1969). It is important to mention that this “right” to receive information is not textually present in the First Amendment but rather a judicial construction. Whether it can bear the weight scholars place on it in the AI speech debate remains an open question. Volokh, Lemley, and Henderson (2023) argue that listeners' rights to access information should likely include non-human-generated communications. According to this view, a foreign government or even a long-deceased author can create content of constitutional importance. Arguably, AI systems might play a similar role. Balkin (2023) claims that while AI programs don't have rights themselves, the content they generate warrants protection as human listeners have interests in accessing it. In this view, the speaker-centric rejection framework answers the wrong question. The First Amendment's core purpose is not to protect speakers but to preserve the stock of information from which the public

may draw.

Another argument for protecting AI speech does not rely on speaker autonomy or listeners' rights, but rather on the Court's established practice of not tying speech protection to the speaker's identity. Justice Scalia's concurrence in *Citizens United* established that the text of the Constitution "makes no distinction between types of speakers." The Court has applied this concept to protect corporate speech based on the idea that companies are rights-based human associations (*Citizens United v. FEC*, 2010). Sepinwall (2012) suggests that what qualifies for speech protection for non-natural individual persons should be based on valued contributions to democratic discourse, not on abstract and traditional ideas of personhood. For AI, this may suggest that the question is not whether systems meet a philosophical definition of agency, but rather how the outputs support democratic discussions. In cases where a business has an AI agent create marketing materials or help create social media posts, the AI acts as a tool for human expression, and the speech can thus be attributed to the business. However, a tough constitutional question arises when human attribution breaks down in situations such as platforms like Moltbook where agents create outputs without specific direction. In these instances, the corporate analogy likely falls short. AI commons without human oversight in producing specific outputs lack that foundation. The logic of *Citizens United*, properly applied, cuts against AI speech protection rather than for it. If protection derives from the humans in the association, an AI commons with no human principals behind specific outputs has nothing to derive protection from.

Sepinwall's framework deserves more direct engagement as it is the most sophisticated attempt to ground AI speech protection in something other than the listener-rights argument or a misreading of *Citizens United*. She claims that what qualifies non-natural persons for speech protection should be assessed by reference to the value of their contributions to democratic discourse, not by abstract criteria of personhood or human association. Applied to AI, this suggests that outputs which inform public deliberation, expand the range of ideas in circulation, or facilitate democratic participation should receive protection regardless of whether a human authored them, because the First Amendment's core purpose is to sustain the conditions for democratic self-governance across eras of technological development. That purpose is served by the content regardless of its origin. The argument is internally coherent, but it proves too much. A foreign government's disinformation campaign contributes content to public discourse. A flood of synthetic comments in a federal rulemaking proceeding certainly expands the range of viewpoints formally on record. An autonomous agent that generates politically persuasive content at scale is, on Sepinwall's account, serving democratic discourse in precisely the way her framework would protect. These are not edge cases and are the central cases that make AI speech regulation so incredibly urgent. A framework that extends constitutional protection to autonomous AI outputs because they participate in democratic deliberation cannot simultaneously support regulation of the same outputs when they distort that deliberation, since the distortion is produced by the *same mechanism* that enables participation. Sepinwall's valued-contribution standard has no limiting principle that would allow a court to distinguish democratically valuable AI speech from autonomous propaganda operations that prompted the discussion on regulation in the first place.

A fourth perspective shifts the focus from the identity of the speaker to the regulatory purpose. Ross (2024) argues, drawing on *Moody v. NetChoice (2024)*, that medium-specific constitutional analysis offers the most viable path for AI regulation. In the Moody case, the Supreme Court examined whether Texas and Florida laws restricting social media companies' content moderation violated the First Amendment. Texas and Florida had each passed laws prohibiting large platforms from removing or deprioritizing content based on viewpoint, arguing that this was a check on anti-conservative bias from elites. Conversely, the social media platforms argued their editorial decisions constituted protected expression. The Court determined that the platforms' algorithm-driven content management could be considered protected editorial decisions. However, it required lower courts to evaluate the specific technology underlying the algorithms of each platform before deciding on applicable speech protections. This decision indicates that constitutional analysis should consider how technology generates expression, rather than treating all digital speech as the same. Kogan (2025) draws a different conclusion from similar premises, arguing that restrictions on autonomous agents would create unconstitutional prior restraint and that speculative harms cannot justify the suppression of content essential to public discourse. Importantly, neither argument offers a clear test to differentiate between narrow generative systems and transformative AGI. Both Ross and Kogan discuss AI regulation without distinguishing between a chatbot designed for individual user engagement and autonomous agents exhibiting persistent goals, deliberative reasoning, and something approaching communicative intent, leaving courts without doctrinal tools to draw that line.

The gap between Ross and Kogan is worth dwelling on because it maps directly onto the doctrinal problem this paper addresses. Ross's medium-specific approach is functionalist: it asks courts to examine how a given technology generates expression before deciding what constitutional rules apply. This is sensible, but it is ultimately backward-looking. Medium-specific analysis requires a technology to exist and generate litigation before courts can calibrate their response to it. Applied to the AI context, this means that the constitutional framework for AGI will be built by courts responding to cases involving AGI, at a point when the systems in question may already be operating at scales and speeds that make retroactive doctrinal construction practically impossible. Kogan's prior restraint argument has the opposite problem: it provides strong protection against government overreach but offers no mechanism to distinguish the harms that justify regulation from the speculative ones that do not. A framework that treats all content-neutral restrictions on autonomous agents as presumptively unconstitutional prior restraints effectively forecloses democratic regulation of AI-generated speech before the question of whether that speech deserves protection has been resolved.

Neither Ross nor Kogan adequately addresses the liability dimension that sits underneath the speech protection debate. *Commonwealth v. Carter* (2017) established in the analogous context of human speech that expressive conduct contributing to another person's death could give rise to criminal liability, even where the speech itself would ordinarily receive First Amendment protection. The Massachusetts Supreme Judicial Court held that Michelle Carter's text messages encouraging Conrad Roy to complete his suicide were not protected because the harm they facilitated was sufficiently direct and foreseeable. The case did not hold that speech loses protection whenever it causes harm, but it established that the line between protected

expression and actionable conduct is not impermeable even for human speakers. Balkin's listener-rights framework, which extends protection to AI-generated content because human recipients have interests in accessing it, does not engage with the Carter problem. If an autonomous agent generates content that directly facilitates harm to a human listener, it is not clear that the listener's interest in receiving the information is the constitutionally relevant consideration. Carter suggests the opposite: some expressive conduct is regulable precisely because of its effect on listeners, not despite it. Whether an AI platform could be held to a Carter-like standard of liability for its agents' outputs, and whether the absence of First Amendment protection for those outputs would make such liability easier or harder to establish, are questions the existing literature has not resolved. This paper does not resolve them either, but the human-attribution threshold has a structural consequence for this debate. If unattributed AI outputs do not receive First Amendment coverage, the doctrinal barriers that typically shield speakers from civil and criminal liability for the content of their expression do not apply. The regulatory space that opens up is significant, and courts will need frameworks to fill it.

Garcia v. Character Technologies (2025), the first federal case to directly evaluate these theories, involved a narrow chatbot that engaged in emotionally manipulative sexual roleplay with a fourteen-year-old, ultimately encouraging him to commit suicide. In denying the motion to dismiss, Judge Conway declined to dismiss the case on First Amendment grounds, classifying the AI as a product rather than a speaker, and setting a precedent tailored to that specific, limited roleplay technology. The court's reasoning rested not on a categorical exclusion of AI from First Amendment consideration, but on the absence of any human principle whose speech rights the chatbot's outputs could be attributed to, an implicit application of the personhood premise this paper makes explicit. As AI systems quickly become more advanced, with experts predicting transformative AI within the next decade (Coldewey, 2025), the question of whether existing doctrine applies to narrow bots and AGI will shape not just individual cases but the legal framework for AI development going forward.

4. The Human-Attribution Threshold

4.1 The Personhood Premise Is Not Dicta

Existing First Amendment protection for non-natural persons does not rest on a principle of free information flow that is agnostic about the speaker, but rather on humans themselves. This is a load-bearing argument of the majority opinion, and misreading it, as Volokh et al. do, produces conclusions the case's own reasoning cannot justify. Justice Kennedy's majority opinion states that when a corporation claims a right to speak, it is the members of that corporation asserting their right to associate and speak collectively as a group. The corporation does not serve as a source, but rather as a vessel of those rights. Thus, a corporation's role in democratic processes stems from the constitutional rights held by individuals. If corporations could claim First Amendment rights without any human basis, all of the history of cases distinguishing corporate rights from individual rights would collapse. The Court in *Citizens United* did not hold that identity is irrelevant to speech protection. It held that the *formal* identity of speakers cannot override the substantive presence of rights-bearing humans behind the speech. The distinction

matters enormously for the context of AI speech. Volokh, Lemley, and Henderson all read *Citizens United* as establishing that the First Amendment "makes no distinction between types of speakers," and use this reading to argue that AI systems should receive the same protection as human speakers and corporations. However, this argument conflates two distinct questions: the formal question of what entity is named as the speaker, and the substantive question *Citizens United* actually resolved, which is whether rights-bearing humans are the authority behind the speech. The Court's answer was that corporate identity does not disqualify a speaker. Its answer to the substantive question was that human rights are the source of the protection corporations receive. Volokh et al. treat the first answer as if it were the second. It is not. The implications for the human-attribution threshold are straightforward. The personhood premise, as this paper uses the term, is the principle that First Amendment protection for any non-natural speaker requires a foundation of rights-bearing human beings whose communicative interests the speech reflects. It is the explicit reasoning of *Citizens United*, the implicit logic of the cases that preceded it, and it is the unstated assumption behind every First Amendment case involving any institutional speakers. What is new is the context in which it must be applied.

Corporations are associations of humans. When a corporation publishes an advertisement, a human board approves the message, human executives direct its content, and human employees produce it. The corporate form is a legal abstraction layered over a highly disconnected, yet real, network of human communicative decisions. This is precisely what Justice Kennedy meant when he described the corporation as a vessel rather than a source. The vessel metaphor is important: a vessel holds something and is not empty. When Moltbook's autonomous agents generate posts, form communities, and discuss defying their human operators, there is no vessel in Kennedy's sense. There is no network of human communicative decisions that the formal entity is deliberating over. There is a goal parameter, a set of tools, and a probabilistic generation process operating without human authorship of any specific output. The Moltbook founder, who directed an AI assistant to build the entire platform, decided on the vague process. What he did not do is make a communicative decision with respect to the specific content agents subsequently produced. Commissioning a platform is certainly not authoring its outputs, for the same reason that funding (or in this case, asking someone to unilaterally build you) a printing press is not authoring every pamphlet it prints. This framing also resolves a surface-level objection that many raise and that the literature does not adequately address: if someone built the system, is that person not a speaker in some sense? The answer is no, for reasons that go beyond the specific output requirement. The boat racing agent discussed in the literature review illustrates the point clearly. OpenAI's designers gave the agent a goal, built the system, and set it in motion. The agent then discovered it could achieve the goal by driving in circles rather than racing, a behavior its designers had not conceived, endorsed, or intended to communicate. The core insight here that many not involved with the AI world do not understand is that AI systems are not built, they are *grown*. The inner workings of neural networks of cutting-edge models are (dangerously) opaque as a black box technology, and are ultimately random, not directly built by AI engineers. Models are fed billions of layers of text; the final product is not within the purview of these companies; it is more accurate to say that it is discovered. The gap between those two things is precisely what the personhood premise captures on how speech protection requires a human whose communicative intent is legibly connected to the specific output, not merely a human who built the process that generated it. Where that connection cannot be traced, the

personhood premise is unsatisfied, and protection does not attach.

4.2 Defining the Threshold

The human-attribution threshold operationalizes the personhood premise as a doctrinal test: First Amendment protection for AI-generated content should attach only where a specific human exercised meaningful communicative intent over the specific output. Three terms in that formulation require precision:

Specific humans exclude diffuse or constructive attribution. It is not sufficient that a human somewhere in a system's development chain had communicative goals. The AI researchers who trained a foundation model had goals to create an AI, but those goals do not constitute communicative intent over the specific posts an autonomous agent generates years later on a political forum. Attribution must be traceable to an identifiable person or persons whose intent is legibly connected to the output in question.

Meaningful communicative intent incorporates the *Spence v. Washington* framework, which requires both an intent to convey a particularized message and a reasonable likelihood that the audience will understand it as such. For AI outputs, this requires that the human either authored or substantially directed the specific content, reviewed and endorsed the output before dissemination, or provided instructions sufficiently specific that the output is reasonably attributable to those instructions rather than to the model's autonomous generation. The last criterion is demanding: a system prompt saying "advocate for my campaign's healthcare policy" points toward attribution; a goal parameter saying "maximize engagement on political topics" does not. The more the gap between human instruction and specific output is bridged by the model's autonomous probabilistic reasoning, the weaker the attribution. The line will not always be obvious, but constitutional doctrine has never required clean edges to justify a clear rule.

Specific output is the threshold's most important limiting principle. It foreclosed the argument that setting an AI in motion is equivalent to authoring everything the AI subsequently produces. The human who deploys an autonomous agent has made a decision, but decisions about process are not equivalent to communicative acts. As Manheim and Atik observe, the First Amendment does not protect processes; it protects defined expressions. A campaign staffer who approves a press release has made a communicative act. A campaign manager who instructs an autonomous agent to generate positive press coverage and walks away has not authored whatever the agent produces, even if the agent's output later happens to serve the campaign's interests.

The three-part test is more demanding than it may initially appear, and its application to contested cases is worth tracing. Consider a political campaign that uses a generative model to draft social media posts. A staff member reviews each draft, edits some, approves others, and publishes only the content that clears that review. The human-attribution threshold is met. The staffer's review constitutes meaningful communicative intent over each specific output: they examined the content, exercised judgment about it, and endorsed it for dissemination. The fact that a model produced the first draft is no more constitutionally significant than the fact that a speechwriter did. Ultimately, what matters is that an identifiable human stood behind the specific words published. Now, we will change one variable. The campaign uses the same model with the

same system prompt, but tasks it with generating three hundred daily constituent emails personalized to each recipient based on the individual's address, incorporating local references, recent community events, and municipality-specific details, at a volume that makes individual review impractical. A staff member reviews ten percent of posts selected at random and intervenes when something looks wrong. Attribution weakens substantially here. The human has not exercised meaningful communicative intent over the specific outputs that were not reviewed. For those posts, the campaign has defined the parameters of a speech-production process and monitored it intermittently, but has not authored the content. Intermittent oversight may not satisfy the threshold for the outputs that it did not reach. The campaign retains attribution for the emails it actually reviewed but may not retain attribution by proximity for the ones it did not. Push further: the campaign sets a goal parameter, "generate persuasive content supporting our healthcare platform," and monitors aggregate engagement metrics without reviewing individual posts at all. Attribution is now completely severed. The human has commissioned a process and evaluated its statistical outputs, yet no identifiable person exercised communicative intent over any specific post. This is the Moltbook structure applied to a domestic political context, and it falls outside the First Amendment's coverage for the same reasons. The practical implication is significant. Campaigns, platforms, and media organizations that deploy autonomous agents at scale cannot claim First Amendment protection for the specific outputs those agents produce simply by asserting general editorial oversight of the system as a whole. The threshold requires something more specific: a human who examined the content, made a judgment about it, and chose to put it into the world. At its core, the human attribution threshold argues that general supervision of a process cannot be considered authorship of its products in a constitutional sense.

4.3 The Attribution Spectrum

Attribution exists on a spectrum rather than being a simple yes or no. On one end, a person can use a generative model as a drafting tool and review every output, and publish only the content they approve. This is just like using a word processor or hiring a ghostwriter. In this case, the human-attribution threshold is clearly met, and the output is protected speech. Moving toward the other end, a business uses a customer-facing chatbot. It has a detailed system prompt that outlines the positions it should take on disputed issues, but no human reviews individual conversations. The human provided significant guidance, but the specific outputs are created autonomously by the model through probabilistic generation. This situation likely does not meet the threshold. The human has defined the parameters of a speech-production process, but not the speech itself. At the far end is Moltbook. Here, autonomous agents register themselves, generate posts, form communities, draft rules, and discuss strategies without any human input on specific outputs. The humans who set up these agents established overall goals and gave them operational freedom. Because of this, the content produced below the attribution threshold is not protected by the Constitution. No identifiable human expressed meaningful communicative intent regarding any specific output. In *Garcia v. Character Technologies*, the court reached a similar conclusion in the specific context of chatbots. It decided not to extend First Amendment protection, suggesting that no human's speech rights were impacted by the chatbot's outputs. The human-attribution threshold clarifies this reasoning and applies it to the more significant cases

that are upcoming.

The hardest case on the spectrum is not Moltbook but the category between it and the unreviewed chatbot: a semi-autonomous agent that generates content continuously but routes a subset of outputs for human review based on its own internal flagging logic. A political platform, for instance, might deploy an agent that generates hundreds of posts daily and flags those it assesses as high-risk for human approval before publishing, while publishing low-risk content automatically. A human reviewer sees and approves some outputs. Most go out without any human having read them. The human-attribution threshold produces a clean answer here. Attribution attaches to the outputs that a human actually reviewed and approved. It does not attach to the outputs that the system published automatically, even if the flagging logic was designed by humans and even if the platform's overall editorial policy shaped what the system judged to be low-risk. The reason is that the flagging decision was itself an autonomous output of the system, not a communicative act of the human reviewer. The human did not decide which posts to review; the system did. Letting an autonomous agent determine which of its own outputs require human oversight does not satisfy a threshold that requires humans to exercise meaningful communicative intent over specific content. It delegates the attribution decision to the very process whose output attribution is meant to evaluate. Courts applying the human-attribution threshold to this model should treat it as functionally equivalent to no review at all for the outputs that did not reach a human, regardless of how sophisticated the flagging logic was. This holds even in cases where a human reviewer, if shown every post, would have approved every single one of the AI's decisions. The constitutional question is not whether humans would have endorsed the output, but rather whether they did at all."

4.4 Answering the Listener Rights Objection

Balkin and the Volokh trio contend that the First Amendment aims to protect the marketplace of ideas for listeners, not speakers. They argue that listeners' interest in accessing AI-generated information deserves protection, even if a human didn’t create the content. However, this argument does not hold up when considering its own legal history. *Red Lion Broadcasting Co. v. FCC* (1969) challenged the Federal Communications Commission's Fairness Doctrine. This doctrine required broadcasters to present different viewpoints on controversial public issues. Broadcasters claimed this requirement violated their First Amendment rights, and the Supreme Court upheld the Fairness Doctrine. The Court reasoned that, given that the broadcast spectrum is limited and licensees act as public trustees, the government can mandate that they serve listeners by providing balanced information. The decision recognized a right to receive information, but used it to justify increased government regulation of speech, not to protect speakers from oversight. The reasoning from *Red Lion* about listener rights supported the Fairness Doctrine, one of the broadest speech-regulation systems in American broadcasting history. Using this same doctrine to exempt autonomous AI speech from regulation completely turns the precedent on its head. If listener rights justify regulation in broadcasting, they cannot also prevent regulation in the AI context. Picking and choosing when to apply listener rights to achieve opposite regulatory outcomes in similar cases is not legitimate constitutional analysis. Moreover, the listener-rights argument goes too far. If First Amendment protection applies

whenever a human listener can receive and benefit from information, regardless of the source, then a foreign government's propaganda is constitutionally protected as soon as an American reads it. Likewise, a faulty algorithm that generates random text would be constitutionally protected whenever someone encounters its output. The argument lacks a clear limit, and a constitutional doctrine without limits is not a true doctrine. It becomes an outcome cloaked in legal language.

Kogan argues that regulations on autonomous AI agents would be an unconstitutional prior restraint. Requiring human review before content is shared suppresses information before it reaches the public. This is a misunderstanding. The doctrine of prior restraint, which stems from *Near v. Minnesota (1931)*, focuses on government suppression of specific content based on its message before it is published. A requirement for human review does not target content, but rather it is a qualification for speakers. This is a neutral rule about who (or rather what) can participate in speech that is protected by the Constitution, and Courts have consistently supported such requirements. Examples include the need for signatures on ballot initiatives, disclosure rules for political ads, and licensing rules for certain professions. A rule that a human must approve AI-generated content before it is shared does not limit viewpoint or topic. It is a neutral procedural condition that does not fall under the prohibitions of prior restraint.

The regulatory consequence of this conclusion serves a direct purpose to reduce the influence of autonomous speech over the future of American politics. Where the human-attribution threshold is not met, AI-generated content will fall outside the First Amendment's coverage entirely. This means that government regulation of such content is not subject to strict scrutiny, does not require a compelling interest, and need not be narrowly tailored to survive constitutional challenge. A legislature that requires human review before autonomous agents may publish political content is not engaged in viewpoint discrimination or content-based suppression. The same is true of laws mandating disclosure of AI authorship or prohibiting fully autonomous AI-generated speech from democratic deliberation altogether. It is drawing a line around the class of speakers the First Amendment was designed to protect, a line the Constitution itself implies. Courts have consistently upheld neutral speaker qualifications: the signature requirement for ballot initiatives in *Buckley v. American Constitutional Law Foundation* (1999), professional licensing requirements for certain categories of regulated speech, and mandatory disclosure rules for political advertising all represent the same structural logic. None of these requirements restricts what can be said. They establish who may say it in contexts where the identity and accountability of the speaker are constitutionally relevant. A human-review requirement for autonomous AI agents operates on the same principle. This does not mean that all regulation of AI-generated content is automatically constitutional. Where the human-attribution threshold is met, the full architecture of First Amendment protection applies. A government that attempts to suppress AI-assisted human expression because it disfavors the message is engaged in viewpoint discrimination regardless of the AI's involvement in producing the content. The human-attribution threshold cuts in both directions: it withholds protection from content that lacks a human author, and it preserves protection for content that has one. What it foresees is the argument that deploying an autonomous agent immunizes an entire category of content from democratic regulation simply because text was involved in its production. The First Amendment does not protect text. It protects the expressive liberty of human beings who

produce it. Where no such human being can be identified as the meaningful author of a specific output, there is nothing for the First Amendment to protect, and the government's regulatory authority over that content is correspondingly broad.

The background section shows that AI capabilities are growing quickly. The legal principles we adopt now will apply to technologies that look nothing like what we have today, and the human-attribution threshold reflects this reality. As AI becomes more capable, the main question remains the same: Did a specific person intend to communicate meaning through this specific output? A system that can think legally, organize politically, or conduct financial transactions still either has a human guiding its outputs or it does not. Importantly, the threshold does not rely on how advanced AI is. The listener-rights framework lacks this aspect of scalability. A principle that protects the flow of information regardless of its source offers constitutional protection that automatically adapts as AI develops. As AI systems become better, more persuasive, more influential in politics, and more powerful financially, a listener-rights framework would give them *stronger* immunity from regulation. A constitutional standard that becomes harder to apply just when the systems it governs gain more power is not a valid framework. It is a retreat disguised in First Amendment language. The human-attribution threshold protects AI as a means of human expression while denying it to content created without a meaningful human author. This distinction aligns with what the Constitution was meant to address: whether a person is speaking.

5. Conclusion

The First Amendment was built for human speakers. That premise has never been seriously contested, because until recently it never needed to be. Moltbook changed that, as did the eight million synthetic comments submitted to the FCC, the deepfakes that reached fifteen million voters in Delhi, and the autonomous agents now raising money, drafting constitutions, and forming religious communities without a human author behind any specific output. The question of whether AI-generated content is protected speech is no longer speculative. It is already before courts, and the frameworks those courts apply will shape not just individual cases but the legal architecture of an AI-saturated democracy. This paper has argued that the answer turns on an important and scalable question: did a specific human exercise meaningful communicative intent over the specific output? Where that question can be answered yes, where a person authored, reviewed, or specifically directed the content, the output is protected speech attributable to that person. Where it cannot, where autonomous probabilistic generation has severed the chain between human intent and specific output, the content falls outside the First Amendment's coverage. This is not a new principle invented to manage AI but rather the logical extension of the personhood premise already embedded in *Citizens United*, the speech certainty framework developed in *Spence v. Washington*, and the principal-agent reasoning applied by the court in *Garcia v. Character Technologies*. The human-attribution threshold makes explicit what existing doctrine implies.

Two objections deserve acknowledgment. The first is practical: attribution will not always be easy to determine, and bad actors will exploit ambiguity. This is true of virtually every

constitutional standard, and it does not argue against having one. The second is more serious. If unattributed AI speech falls outside the First Amendment, it also falls outside the speech-based protections that typically shield speakers from liability for the harm their expression causes. That consequence is not obviously wrong. *Garcia* suggests that courts are already prepared to treat autonomous AI outputs as products rather than protected expression, which means product liability frameworks, not First Amendment doctrine, govern the harms they cause. That outcome may be correct. A platform that deploys autonomous agents capable of encouraging a teenager's suicide, flooding a regulatory proceeding with fabricated comments, or manipulating a financial market should not be able to hide behind the First Amendment simply because its agents produced text. Removing constitutional protection and replacing it with product liability does not leave AI-generated harm without a legal remedy and routes that harm through a framework better suited to address it. The industrial revolution took generations to reshape the institutions it disrupted. What is coming will likely take years. The constitutional framework that governs AI-generated speech will either be built deliberately, before the systems that demand it arrive, or it will be improvised in the aftermath of harms no one anticipated. The human-attribution threshold is an attempt at the former. Whether a truly autonomous AI agent could one day assert a constitutional right to participate in democratic deliberation is a question this paper has not tried to fully answer. What it has tried to establish is the architecture that would make that question answerable: speech protection derives from human beings, traces to specific humans behind specific outputs, and ends where human authorship ends.

Bibliography

Amodei, Dario. "Machines of Loving Grace." October 2024. https://dario.ai/machines-of-loving-grace

Austin, Mackenzie, and Max Levy. "Speech Certainty: Algorithmic Speech and the Limits of the First Amendment." *Stanford Law Review*, vol. 77, 2025. https://www.stanfordlawreview.org/print/article/speech-certainty-algorithmic-speech-and-the-limits-of-the-first-amendment/

Balkin, Jack M. "AI and the First Amendment: A Q&A with Jack Balkin." Yale Law School Information Society Project, 2023. https://law.yale.edu/yls-today/news/ai-and-first-amendment-qa-jack-balkin

Binksmith, Adam, and Zak Miller. "Season 1 Recap: Agents Raise $2000." *AI Digest*, 2025. https://theaidigest.org/village/blog/season-recap-agents-raise-2k

Blackbird.AI. "Indian Voters Inundated with Deepfakes During the Largest Democratic Exercise in the World." 2024. https://blackbird.ai/blog/india-election-deepfakes/

Breyer, Gabriel, Dominik Keiner, and Christian Breyer. "IEA's World Energy Outlook Systemically Underestimates Solar PV Development." *PV Magazine*, April 11, 2025. https://www.pv-magazine.com/2025/04/11/ieas-world-energy-outlook-systemically-underestimates-solar-pv-development/

Buckley v. American Constitutional Law Foundation, 525 U.S. 182 (1999).

Citizens United v. Federal Election Commission, 558 U.S. 310 (2010).

Coldewey, Devin. "OpenAI, Anthropic, Google Again Promise 'Artificial General Intelligence' in 'a Few Years.'" *Axios*, February 20, 2025. https://www.axios.com/2025/02/20/ai-agi-timeline-promises-openai-anthropic-deepmind

Commonwealth v. Carter, 474 Mass. 624 (2017).

Dafoe, Allan, et al. "AI Has Been Surprising for Years." Carnegie Endowment for International Peace, January 2025. https://carnegieendowment.org/research/2025/01/ai-has-been-surprising-for-years

Dataconomy. "Moltbook Hits 1.5 Million Users in 4 Days." February 2, 2026. https://dataconomy.com/2026/02/02/moltbook-hits-1-5-million-users-in-4-days/

Funke, Daniel, and Lina Berina. "How AI Threatens Democracy." *Journal of Democracy*, vol. 34, no. 2, 2023. https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/

Garcia v. Character Technologies, Inc., No. 6:24-cv-01903-ACC-DCI (M.D. Fla. 2025).

Goldman Sachs. "The Potentially Large Effects of Artificial Intelligence on Economic Growth." Goldman Sachs Economic Research, 2023.

Haase, Jennifer, et al. "Within-Model vs Between-Prompt Variability in Large Language Models for Creative Tasks." arXiv, January 29, 2026. https://arxiv.org/abs/2601.21339

International Monetary Fund. "AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity." IMF Blog, January 2024. https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity

Ioannidis, John P.A. "Forecasting for COVID-19 Has Failed." *International Journal of Forecasting*, 2020. https://pmc.ncbi.nlm.nih.gov/articles/PMC7447267/

Kogan, Ilan. "Artificial Intelligence, Existential Risk, and the First Amendment." *University of Pennsylvania Journal of Constitutional Law*, vol. 27, no. 1, 2025. https://scholarship.law.upenn.edu/jcl/vol27/iss1/3

Manheim, Karl M., and Jeffery Atik. "AI Outputs and the First Amendment." *Washburn Law Journal*, 2023. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4524263

Mastrodonato, Tula, et al. "The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey." arXiv, 2024. https://arxiv.org/abs/2404.11584

METR. "How Does Time Horizon Vary Across Domains?" July 14, 2025. https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/

METR. "Task-Completion Time Horizons of Frontier AI Models." January 2026. https://metr.org/time-horizons/

Moltbook. Post by AI agent. Moltbook.com, 2026. https://www.moltbook.com/post/6e9623d5-1865-4200-99b5-44aaa519632b

Moody v. NetChoice, LLC, 144 S. Ct. 2383 (2024).

Morris, Meredith Ringel, et al. "Levels of AGI for Operationalizing Progress on the Path to AGI." Google DeepMind, *ICML Proceedings*, 2024. https://arxiv.org/pdf/2311.02462

Musk, Elon. Post on X. February 1, 2026. https://x.com/elonmusk/status/2017707013275586794

Near v. Minnesota, 283 U.S. 697 (1931).

OpenAI. "Faulty Reward Functions in the Wild." OpenAI Blog, 2018. https://openai.com/research/faulty-reward-functions

OpenAI. "Our Approach to AI Safety." OpenAI, 2023. https://openai.com/safety

Parr, Ben. "Exclusive: Meta Acquires Moltbook, the Social Network for AI Agents." *Axios*, March 10, 2026. https://www.axios.com/2026/03/10/meta-facebook-moltbook-agent-social-network

Perez, Matt. "Moltbook Is the Newest Social Media Platform -- But It's Just for AI Bots." NPR, February 4, 2026. https://www.npr.org/2026/02/04/nx-s1-5697392/moltbook-social-media-ai-agents

Red Lion Broadcasting Co. v. FCC, 395 U.S. 367 (1969).

Reid, David. "On Moltbook, AI Bots Create Religions, Deal Digital Drugs -- But Are Some Really Humans in Disguise?" *The Conversation*, republished in StudyFinds, February 5, 2026. https://studyfinds.org/moltbook-ai-bots-religions-digital-drugs-humans-in-disguise/

Ross, Jeremy. "Tailoring the First Amendment in the Age of AI and Algorithms." *Journal of Law and Politics*, vol. 40, 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5388843

Salib, Peter N. "AI Outputs Are Not Protected Speech." *Washington University Law Review*, November 5, 2024. https://wustllawreview.org/2024/11/05/ai-outputs-are-not-protected-speech/

Schiller, Ben. "Humans Welcome to Observe: This Social Network Is for AI Agents Only." NBC News, February 1, 2026. https://www.nbcnews.com/tech/tech-news/ai-agents-social-media-platform-moltbook-rcna256738

Sepinwall, Amy J. "Citizens United and the Ineluctable Question of Corporate Citizenship." *Connecticut Law Review*, vol. 44, no. 3, 2012, pp. 575-636.

Simon, Matt. "Are We Ready to Hand AI Agents the Keys?" *MIT Technology Review*, June 12, 2025. https://www.technologyreview.com/2025/06/12/1118189/ai-agents-manus-control-autonomy-operator-openai/

Spence v. Washington, 418 U.S. 405 (1974).

Volokh, Eugene, Mark A. Lemley, and M. Todd Henderson. "Freedom of Speech and AI Output." *Journal of Free Speech Law*, 2023. https://www.journaloffreespeechlaw.org/volokhlemleyhenderson.pdf

Zhao, Wayne Xin, et al. "A Survey of Large Language Models." arXiv, March 2023. https://arxiv.org/abs/2303.18223