Stanford Law Review

Volume 77 January 2025

ARTICLE

Speech Certainty: Algorithmic Speech and the Limits of the First Amendment

Mackenzie Austin & Max Levy*

Abstract. Machine learning algorithms increasingly mediate our public discourse—from search engines to social media platforms to artificial intelligence companies. As their influence on online speech swells, so do questions of whether and how the First Amendment may apply to their output. A growing chorus of scholars has expressed doubt over whether the output of machine learning algorithms is truly speech within the meaning of the First Amendment, but none have suggested a workable way to cleanly draw the line between speech and non-speech. This Article proposes a way to successfully draw that line based on a principle that we call "speech certainty"—the basic idea that speech is only speech if the speaker knows what he said when he said it. This idea is rooted in the text, history, and purpose of the First Amendment, and built into modern speech doctrines of editorial discretion and expressive conduct. If this bedrock principle has been overlooked, it is because, until now, all speech has been imbued with speech certainty. Articulating its existence was never necessary. But machine learning has changed that. Unlike traditional code, a close look at how machine learning algorithms work reveals that the programmers who create them can never be certain of their output. Because that output lacks speech certainty, it's not the programmer's speech. Accordingly, this Article contends that the output of machine learning algorithms isn't entitled to First Amendment protection. It reveals that the question of how an algorithm works is constitutionally significant. With the Supreme Court in *Moody v. NetChoice* demanding further inquiry into what constitutes protected expressive activity for social media platforms, that question can no longer be ignored. By failing to distinguish between traditional and machine learning algorithms, we risk sleepwalking into a radical departure from centuries of First Amendment jurisprudence. Protection for the output of machine learning algorithms would, for the first time in the Constitution's history,

* Mackenzie Austin is a practicing attorney in California. She served as a Law Clerk for the U.S. Court of Appeals for the Eleventh Circuit. J.D., Stanford Law School, 2022. Max Levy is a practicing attorney in California. J.D., Stanford Law School, 2022. We're grateful to Eugene Volokh and the editors of the Journal of Free Speech Law, Paul Grewal, Peter Henderson, Evelyn Douek, and the editors of the *Stanford Law Review* for their feedback and comments, and to Rick Rubin for the fateful nudge to put ideas to paper. Although *Stanford Law Review* convention suggests use of a single set of pronouns throughout an Article, we alternate usage between he/him and she/her across top-level Parts of the Article (e.g., Parts I, II, III, etc.) to reflect the gender of both authors.

protect speech that a speaker does not know he has said. Speech certainty provides a novel and principled approach to conceptualizing machine learning algorithms under existing First Amendment jurisprudence.

Table of Contents

Introduction

When does something that isn't yet speech—an intuition, a spark of imagination, an embryonic thought—become speech, and, as a result of that becoming, earn the protection of the First Amendment? This is the rare case where inquiry confirms intuition. An idea becomes speech when it's spoken by a speaker: words actually written, a speech actually given, brushstrokes actually painted. The concept is so plain that it has hardly merited any scrutiny. In this Article, however, we argue that this concept—which we call the principle of "speech certainty"—defines the limits of what constitutes speech under the First Amendment, with important implications for our online public discourse.

The idea is simple: Speech is characterized by speech certainty when the speech can be identified with certainty by the speaker at the moment it is spoken. An audience might misunderstand what the speaker meant, or the speaker may have chosen her words poorly, but those words—the ones that left the speaker's mouth—constitute her speech because the speaker knew for certain what she said when she said it. And because until recently no other character of speech has existed, the First Amendment has only ever protected such speech. The principle of speech certainty is so inherent to the concept of speech that articulating its existence was never necessary.

But that has now changed with the emergence of machine learning algorithms, the outputs of which are claimed as the speech of their creators.[1] Unlike traditional algorithms, in which a programmer dictates rules for the algorithm to follow in perfectly predictable ways, machine learning algorithms write their own rules.[2] And these rules, without exception, calculate probabilities to make predictions.[3] Based on the combination of words in a post published by a particular user, for example, what is the

1. Brief for Respondents at 23, Moody v. NetChoice, LLC, 144 S. Ct. 2383 (2024) (No. 22-277) (arguing that social media platforms "engage[] in speech when disseminating 'curated compilations of speech' created by others"); *id.* at 39 (arguing that social media platforms' acts of editorial discretion "are often determined by proprietary algorithms"); Brief for Petitioners at 27, Moody v. NetChoice, LLC, 144 S. Ct. 2383 (2024) (No. 22-555) (arguing social media platforms' editorial decisions are protected even when executed via algorithm).

2. PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 6 (2015) ("With machine learning, computers write their own programs, so we don't have to."); Cade Metz, *AI Is Transforming Google Search. The Rest of the Web Is Next*, WIRED (Feb. 4, 2016, 7:00 AM), https://perma.cc/C3CY-JFH4 ("By building [machine] learning systems, we don't have to write these rules anymore.").

3. In this Article we focus on a subset of machine learning: supervised machine learning algorithms that rely on a mathematical process called gradient descent to generate probabilities. *See infra* Part III.A.

likelihood that the post includes hate speech? Or, based on the way the pixels in an image are arranged, that the image includes nudity? The nature of these algorithms, however, is that their predictions always leave room for doubt. As a result, they can never be 100% accurate in their output.[4] Neither, then, can their programmers be certain of the contents of that output.[5]

Thus, when online platforms rely on machine learning algorithms to rank, recommend, and remove content, they can never be certain that the content published by their algorithms will align with what they intended to publish.[6] In fact, because the algorithm will always be wrong at least some of the time, it is guaranteed that the algorithm will publish precisely what the platforms intended not to publish.[7] A platform that prohibits hate speech and enforces that prohibition with a machine learning algorithm will inevitably publish hate speech.[8] But because it has outsourced enforcement to an algorithm that writes its own rules, the platform cannot know when, where, or even why that hate speech will appear on its platform.[9] Speech *uncertainty* is as inherent to the use of machine learning algorithms as speech certainty is to traditional speech.

Over the last decade, these machine learning algorithms have come to mediate public discourse—from search engines to social media platforms to,

4. *See* Mike Loukides, *The Machine Learning Paradox*, O'REILLY (June 1, 2017), https://perma.cc/TX4L-9B6Z ("It's also going to be imperfect. How imperfect? That depends on the application. 90-95% accuracy is achievable in many applications, maybe even 99%, but never 100%. That doesn't mean machine learning applications aren't useful. It does mean we have to be aware that machine learning is never going to be a 100% solution . . . .").

5. *Infra* Part IV.B.3.a.

6. For discussion on platforms' use of machine learning algorithms, see Tarleton Gillespie, *Do Not Recommend? Reduction as a Form of Content Moderation*, SOC. MEDIA & SOC'Y, July-Sept. 2022, at 1; Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, NEW AMERICA, https://perma.cc/5QZ6-X9XD (archived Oct. 20, 2024) ("[M]any companies have developed or adopted automated tools to enhance their content moderation practices, many of which are fueled by artificial intelligence and machine learning."); Mike Ananny, *Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance*, KNIGHT FIRST AMEND. INST. (Aug. 21, 2019), https://perma.cc/3R33-A7DF (explaining "the probabilistic logics [platforms] use to govern expression").

7. *Infra* Part III.B.2.f.

8. *See, e.g.*, Complaint ¶¶ 7-13, X Corp. v. Media Matters for Am., No. 23-cv-1175 (N.D. Tex. Nov. 20, 2023) (illustrating that X's algorithms cannot completely prevent unwanted content from appearing in users' feeds); *see also* Evelyn Douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 764 (2021) ("[A] probabilistic conception of online speech acknowledges that enforcement of the rules . . . will never be perfect . . . .").

9. *Infra* Part III.B.2.c.

more recently, artificial intelligence companies.[10] These algorithms shape what we see (and don't see) across wide swathes of the internet, giving rise to a palpable and well-chronicled anxiety across the political spectrum.[11] Of course, anxiety about undue control over the public discourse is something of an American tradition—be it targeted at newspapers, broadcast companies, or, most recently, internet platforms.[12] This Article posits that today's concerns are different. Whether articulated in terms of "surveillance capitalism" or "the tyranny of Big Tech," today's concerns are not merely about the outsized influence of a speaker,[13] nor are they a moral panic concerning a new medium for speech. They arise instead from a new breed of "speech" altogether—the output of machine learning algorithms. Through their operation on social media and search platforms, these algorithms provide each of us with our own window into the world. And these windows are, as a technical matter, opaque.[14] Nobody—not even the companies that create the algorithms—fully understands why they make the decisions they do.[15] In response, legislatures the world over have lurched to exert some measure of control over this new force in our public discourse.[16] Such efforts in the United States, however, raise

10. *See, e.g.*, *supra* note 6; Cristos Goodrow, *On YouTube's Recommendation System*, YOUTUBE OFF. BLOG (Sept. 15, 2021), https://perma.cc/3P5P-6QVQ; *OpenAI's Technology Explained*, OPENAI (Oct. 11, 2023), https://perma.cc/3MK2-6JB2.

11. Emily A. Vogels, Andrew Perrin & Monica Anderson, *Most Americans Think Social Media Sites Censor Political Viewpoints*, PEW RSCH. CTR. (Aug. 19, 2020), https://perma.cc/77JJ-6USP.

12. *See, e.g.*, FCC v. Pottsville Broad. Co., 309 U.S. 134, 137 (1940) (noting "the spur of a widespread fear that in the absence of governmental control the public interest might be subordinated to monopolistic domination in the broadcasting field"); Mia. Herald Publ'g. Co. v. Tornillo, 418 U.S. 241, 249 (1974) (describing "the dominant features of a press that has become noncompetitive and enormously powerful and influential in its capacity to manipulate popular opinion and change the course of events"); Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 661 (1994) (highlighting "the bottleneck monopoly power exercised by cable operators and the dangers this power poses to the viability of broadcast television").

13. *See generally* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (2019) (advancing a prominent critique of the increasing influence of algorithms on society); JOSH HAWLEY, THE TYRANNY OF BIG TECH (2021) (same).

14. W. Nicholson Price II & Arti K. Rai, *Clearing Opacity Through Machine Learning*, 106 IOWA L. REV. 775, 778 (2021) (noting that machine learning "models are often opaque").

15. *See* Price, *supra* note 14, at 778-79 (describing difficulty in understanding machine learning and artificial intelligence "black boxes"); Steven Levy, *AI Is a Black Box. Anthropic Figured Out a Way to Look Inside*, WIRED (May 21, 2024, 11:00 AM), https://perma.cc/VBP6-RUF6 ("We have these systems, we don't know what's going on . . . .").

16. Moody v. Netchoice, LLC, 144 S. Ct. 2383, 2393 (2024) ("[T]he questions of whether, when, and how to regulate online entities, and in particular the social-media giants, are understandably on the front-burner of many legislatures and agencies."); *see also The*

footnote continued on next page

controversial questions of whether and how the First Amendment applies to the output of these algorithms.[17] The Supreme Court's decision in *Moody v. NetChoice* assures that these questions will soon have their answers.[18] Where the Court will land, however, is far from a foregone conclusion. While it ventured that "the editorial judgments influencing the content of [platforms'] feeds are . . . protected expressive activity,"[19] it put a stark asterisk on that conclusion: It was based only on the existing, undeveloped record.[20] The immediate contribution of this Article is to clear a path for that record's development in the lower courts. If followed, it will show in fact what this Article articulates in theory—that the output of online platforms' machine learning models does not qualify for First Amendment protection under the jurisprudence articulated by the *Moody* Court.

Thus, with the First Amendment status of online speech in flux, this Article also raises an alarm. If we fail to recognize the paradigm shift ushered in by machine learning, an outdated understanding of algorithmic speech threatens to lead us into a radical departure from centuries of First Amendment jurisprudence. By protecting the output of machine learning, the Constitution would, for the first time in its history, protect speech that a speaker does not know she has said. Perhaps that departure from First Amendment jurisprudence is the path we should take. Or, perhaps, as we and others argue, it is not. But whatever decision we make, it cannot be justified by simply extending the logic of First Amendment protection for the output of traditional code to the output of machine learning algorithms.[21]

In this Article, we make two arguments. First, that the principle of speech certainty defines the limits of the First Amendment. And second, because machine learning algorithms run afoul of this principle, their output is not

---

Digital Services Act, EUR. COMM'N, https://perma.cc/6DX7-SE3K (archived Oct. 20, 2024) (stating the Digital Service Act's "main goal is to prevent illegal and harmful activities online and the spread of disinformation").

17. *See* Daphne Keller, *Amplification and Its Discontents*, KNIGHT FIRST AMEND. INST. (June 8, 2021), https://perma.cc/PF6H-GKC8 (cataloging the First Amendment challenges facing regulatory models targeting algorithmic amplification).

18. *Moody*, 144 S. Ct. at 2409.

19. *Id.*

20. *Id.* (emphasizing conclusion is based "on the current record"); *id.* at 2393 ("*[T]he current record* suggests that some platforms, in at least some functions, are indeed engaged in expression." (emphasis added)); *id.* at 2403 ("[T]he record is *incomplete* even as to the major social-media platforms' main feeds . . . ." (emphasis added)); *see also id.* at 2412 (Jackson, J., concurring in part) ("[F]urther factual development may be necessary before either of today's challenges can be fully and fairly addressed."); *id.* at 2422 (Alito, J., concurring) (emphasizing "the incompleteness of this record").

21. *See id.* at 2410 (Barrett, J., concurring) ("[T]he way platforms use this sort of technology might have constitutional significance.").

speech within the meaning of the First Amendment and thus falls outside its protection.

We reach these conclusions by applying the most widely-accepted modes of constitutional interpretation—text, history, precedent, and purpose—to show that they compel the adoption of the principle of speech certainty.[22] In Part I, we establish that the text, history, and purpose of the First Amendment all support the principle of speech certainty. In Part II, we demonstrate that speech certainty has always been a first-order assumption underlying First Amendment jurisprudence, including the modern doctrines of editorial discretion and expressive conduct. Setting the First Amendment aside, Part III explains how machine learning algorithms work, distinguishing them from traditional algorithms and illustrating the uncertainty intrinsic to the technology. And finally, Part IV applies the technical discussion to the doctrine, revealing that the output of machine learning algorithms lacks speech certainty and is not protected by the First Amendment. Finally, Part V briefly explores the regulatory implications of a First Amendment cabined by the principle of speech certainty.

Ultimately, this Article builds on Lawrence Lessig's and others' intuition that "[a]t some point along the continuum between your first program, 'Hello world!' and [artificial intelligence], the speech of machines crosses over from speech properly attributable to the coders to speech no longer attributable to the coders."[23] We believe that point is defined by speech certainty.

I. The First Amendment and the Principle of Speech Certainty

The text sits patiently atop the Bill of Rights: "Congress shall make no law . . . abridging the freedom of speech, or of the press . . . ."[24] Over the last century, constitutional scholars have meticulously probed these words in their quest for an original understanding of the Constitution.[25] Yet these words,

22. BRANDON J. MURRILL, CONG. RSCH. SERV., R45129, MODES OF CONSTITUTIONAL INTERPRETATION 3 (2018), https://perma.cc/99FX-M4YZ (identifying textualism, original meaning, and judicial precedent as three prominent modes of constitutional interpretation); Alexander Tsesis, *Free Speech Constitutionalism*, 2015 U. ILL. L. REV. 1015, 1016 (2015) (recognizing the ongoing debates over the First Amendment's purposes).

23. Lawrence Lessig, *The First Amendment Does Not Protect Replicants*, *in* SOCIAL MEDIA FREEDOM OF SPEECH AND THE FUTURE OF OUR DEMOCRACY 273, 276 (Lee C. Bollinger & Geoffrey R. Stone eds., 2022).

24. U.S. CONST. amend. I.

25. Genevieve Lakier, *The Non-First Amendment Law of Freedom of Speech*, 134 HARV. L. REV. 2299, 2303 (2021) (noting that the First Amendment tradition "began to emerge in its modern form only in the early decades of the twentieth century").

more than any others, elude their grasp.[26] Fourteen deceptively simple words, promising the full power of the Constitution to whomever wrests original meaning from the text. The First Amendment—the originalist's sword in the stone.[27]

Like King Arthur's Excalibur, however, the original meaning of the First Amendment is a myth.[28] What the framers and ratifiers of the Constitution meant by "the freedom of speech" has been subject to a century of scholarly disagreement with no resolution in sight.[29] And we make no attempt to resolve it here.

Rather, in this Part, we zoom in to ask a simpler question: What did the Framers mean by "speech"? By that, we don't mean what *forms* of speech—the spoken, written or printed word, for example—were meant to be included.[30] Instead, we seek to understand whether the original public meaning of the word can tell us when something that isn't yet speech—an unexpressed thought or idea—becomes speech, and, as a result of that becoming, earns the protection of the First Amendment.

---

26. LEONARD W. LEVY, LEGACY OF SUPPRESSION: FREEDOM OF SPEECH AND PRESS IN EARLY AMERICAN HISTORY 4 (1960) ("The meaning of no other clause of the Bill of Rights at the time of its framing and ratification has been so obscure to us.").

27. 1 RODNEY A. SMOLLA, SMOLLA AND NIMMER ON FREEDOM OF SPEECH § 1:11 (2016) ("One can keep going round and round on the original meaning of the First Amendment, but no clear, consistent vision of what the framers meant by freedom of speech will ever emerge."); David Lat, *Justice Scalia, Originalism, Free Speech and the First Amendment*, ABOVE THE LAW (Nov. 22, 2016, 6:58 PM), https://perma.cc/X7PF-EBQH ("Free speech has been kind of a desert when it comes to originalism." (quoting Michael McConnell)); *see also* Jack M. Balkin, *Nine Perspectives on* Living Originalism, 2012 U. ILL. L. REV. 815, 837 (2012) ("The abstract language of the First Amendment left unresolved differing views about the meaning of freedom of speech and press; these disputes would break out into the open later on in the 1790s . . . ."); Jud Campbell, *What Did the First Amendment Originally Mean?*, 31 RICHMOND L. MAG., Summer 2018, at 19, 20 ("If the founders couldn't even agree among themselves about that type of law, then surely looking for the First Amendment's 'original meaning' is like searching for the Holy Grail.").

28. *See* Caroline Mala Corbin, *Free Speech Originalism: Unconstraining in Theory and Opportunistic in Practice*, 92 GEO. WASH. L. REV. 633, 636 (2022) ("[T]he theory of originalism applied to freedom of expression is especially ill-advised because we can confidently conclude little about the original meaning of the First Amendment."). *See generally supra* notes 26-27 (collecting sources).

29. Robert H. Bork, *Neutral Principles and Some First Amendment Problems*, 47 IND. L.J. 1, 20 (1971) ("The law has settled upon no tenable, internally consistent theory of the scope of the constitutional guarantee of free speech."); Lat, *supra* note 27.

30. Scholars generally agree that "speech" includes spoken, written, printed, or symbolic speech. *See, e.g.*, Eugene Volokh, *Symbolic Expression and the Original Meaning of the First Amendment*, 97 GEO. L.J. 1057, 1080 (2009). We also agree with Judge Bork that "[n]o one, not the most obsessed absolutist," holds the position that "speech" is limited to verbal communication. Bork, *supra* note 29, at 21.

Here, we show that while the original meaning of "*the freedom of speech*" may not be ascertainable, the original meaning of "*speech*" is. Specifically, at the time of the Founding, the only speech that was granted the First Amendment's protection was the speech that the speaker knew he communicated when he communicated it. Founding-era dictionaries reveal a sharp distinction between pre-speech thoughts and those that have manifested themselves as speech. And to those who lived in the Founding era—at a time when only oral, written, or printed communications existed—the notion that speech could even lack speech certainty was simply unfathomable. These findings reveal the outer limits of the Framers' conception of the First Amendment: Speech is protected only when the speaker knows with certainty what he says at the moment he says it.[31]

A. Text: Founding-Era Definitions of "Speech"

Dictionaries offer our first clues that "speech" includes an inherent principle of speech certainty. Within the nine most credible Founding-era dictionaries, the definitions of "speech" are surprisingly undifferentiated.[32] Together, they articulate nine definitions, eight of which are plausibly relevant to the First Amendment: Speech is (1) an "articulate utterance," (2) "expressing thoughts" or "ideas," (3) "any thing spoken," (4) "talk," (5) "oration," (6) "speaking," (7) a "declaration of thoughts," and/or (8) a "conveyance from one man's mind to another."[33]

---

31. Whether the original public meaning of "speech" alone *compels* the adoption of the principle of speech certainty will depend on the reader's feelings about originalism as a method of constitutional interpretation. The authors take no position on originalism. We merely show that the original public meaning of "speech" is both ascertainable and consistent with the principle of speech certainty. In conjunction with this Article's discussion of how speech certainty also comports with the First Amendment's purpose (Part I.C), relevant precedent (Part II), and contemporary considerations (Parts IV and V), we believe this Article compels the adoption of the principle of speech certainty, whatever one's preferred mode of constitutional interpretation may be.

32. Gregory E. Maggs, *A Concise Guide to Using Dictionaries from the Founding Era to Determine the Original Meaning of the Constitution*, 82 GEO. WASH. L. REV. 358, 382-90 (2014) (identifying the nine most "commonly available and regularly cited" Founding-era dictionaries).

33. *See Speech*, 2 JOHN ASH, THE NEW AND COMPLETE DICTIONARY OF THE ENGLISH LANGUAGE (London, Edward Dilly, Charles Dilly & R. Baldwin 1775) ("articulate utterance," "expressing thoughts," "talk," "any thing spoken"); *Speech*, 2 NATHAN BAILEY, THE NEW UNIVERSAL ETYMOLOGICAL ENGLISH DICTIONARY (London, T. Waller, 4th ed. 1756) ("conveyance of one man's mind to another"); *Speech*, JAMES BARCLAY, A COMPLETE AND UNIVERSAL ENGLISH DICTIONARY (London, J.F. & C. Rivington et al. 1792) ("expressing our thoughts or ideas"); *Speech*, THOMAS DYCHE & WILLIAM PARDON, A NEW GENERAL ENGLISH DICTIONARY (London, Toplis, Bunney & J. Mozley, 18th ed. 1781) ("conveyance of one man's mind to another"); *Speech*, SAMUEL JOHNSON, A DICTIONARY OF THE ENGLISH LANGUAGE (London, J.F. & C. Rivington et al., 10th ed.
footnote continued on next page

Of these, by far the most common—and the one listed as the primary definition in seven of the nine dictionaries—is some variation of the definition in Samuel Johnson's authoritative edition: "The power of articulate utterance; the power of expressing thoughts by words."[34] Given its prominence in individual dictionaries and across the set of Founding-era dictionaries, this definition likely captures how most people would have understood the term "speech" as used in the First Amendment.[35]

The two main elements of Johnson's definition—"articulate utterance" and "expressing thoughts"—make clear what speech actually is: the external manifestation of something that previously existed only in the speaker's mind.

- *"Articulate Utterance."* Speech, in its most literal definition, is an "articulate utterance." According to Founding-era dictionaries, this means that one's "manner of speaking," "pronunciation," or "vocal expression, emission from the mouth" ("utterance")[36] is delivered in a manner "distinct, very plain, and easy to be heard"[37] "so as to form words"[38] ("articulate"). Thus, under this definition, speech relies on a

1792) ("articulate utterance," "expressing thoughts," "oration"); *Speech*, WILLIAM PERRY, THE ROYAL STANDARD ENGLISH DICTIONARY (Worcester, 1788) ("talk," "articulate utterance"); *Speech,* 2 THOMAS SHERIDAN, A COMPLETE DICTIONARY OF THE ENGLISH LANGUAGE (London, Charles Dilly, 3d ed. 1790) ("articulate utterance," "expressing thoughts," "any thing spoken," "talk"); *Speech*, JOHN WALKER, A CRITICAL PRONOUNCING DICTIONARY AND EXPOSITOR OF THE ENGLISH LANGUAGE (London, G.G.J. & J. Robinson & T. Cadell, 1791) ("oration," "talk," "any thing spoken," "articulate utterance," "expressing thoughts"); *Speech*, 2 NOAH WEBSTER, AN AMERICAN DICTIONARY OF THE ENGLISH LANGUAGE (New York, S. Converse 1828) ("expressing thoughts," "talk," "any declaration of thoughts").

34. *Speech*, JOHNSON, *supra* note 33; Maggs, *supra* note 32, at 359 (identifying Johnson's dictionary as "one of the most authoritative eighteenth-century dictionaries"). The remaining two dictionaries offer only one definition of speech and define it by the same concept, though perhaps more esoterically. *Speech*, BAILEY, *supra* note 33 ("that admirable conveyance of one man's mind to another"); *Speech*, DYCHE & PARDON, *supra* note 33 ("that wonderful conveyance of one man's mind to another").

35. *See also* Maggs, *supra* note 32, at 382 ("[T]he more dictionaries consulted, the more persuasive and reliable is the evidence found.").

36. *Utterance*, JOHNSON, *supra* note 33; *Utterance*, WEBSTER, *supra* note 33 ("pronunciation; manner of speaking . . . emission from the mouth; vocal expression"); *Utterance*, SHERIDAN, *supra* note 33 ("pronunciation, manner of speaking . . . vocal expression, emission from the mouth"); *Utterance*, ASH, *supra* note 33 ("[p]ronunciation, vocal expression"); *Utterance*, PERRY, *supra* note 33 ("pronunciation"); *see also Utterance*, DYCHE & PARDON, *supra* note 33 ("Speech, or the way or mode of speaking"); *Utterance*, BARCLAY, *supra* note 33 ("manner or power of speaking").

37. *Articulate*, DYCHE & PARDON, *supra* note 33; *see also Articulate*, JOHNSON, *supra* note 33 ("[d]istinct"); *Articulate*, ASH, *supra* note 33 ("distinct"); *Articulate,* SHERIDAN, *supra* note 33 ("[d]istinct"); *Articulate*, WALKER, *supra* note 33 ("[d]istinct").

38. *Articulate*, BARCLAY, *supra* note 33; *see also Articulate*, WEBSTER, *supra* note 33 ("articulation of the organs of speech"); *cf. Articulate*, PERRY, *supra* note 33 ("distinct in speech").

person actually speaking orally.[39] An unverbalized thought, in other words, is not speech.

- *"Expressing Thoughts."* More broadly, Johnson also defines speech as "expressing thoughts."[40] In every Founding-era dictionary, "to express" means "to represent in words," or "by any of the imitative arts," "to exhibit in language," "to speak" or "to show or make known in any manner."[41] Thus, to express thoughts means to translate "thoughts"—"that which the mind thinks"[42]—from the private realm of the mind into some communicable form. An unexpressed thought, then, is also not speech.

Anchored by these two definitions, the primary Founding-era definition of speech reveals a clear distinction between thoughts and speech, between the "act[s] or operation[s] of the mind"[43] and their outward expression to others. Prior to their utterance or expression, thoughts have not yet become speech. But once uttered or expressed—verbally, written, or in any of the "imitative arts"[44]—those thoughts are transformed into speech. As Johnson wrote in an illuminating explanation accompanying the definition of speech:

> Though our ideas are first acquired by various sensations and reflections, yet we convey them to each other by *the means of certain sounds, or written marks*, which we call words; and a great part of our knowledge is both obtained and communicated by *these means, which are called speech.*[45]

---

39. Although these Founding-era dictionaries frame the term "utterance" exclusively in terms of verbal expression, no scholar—even ardent originalists—understands "speech" within the First Amendment to be so limited. *See* Bork, *supra* note 29, at 21.

40. *Speech*, JOHNSON, *supra* note 33.

41. *See Express*, ASH, *supra* note 33 ("to represent in words," "to shew [sic] or make known in any manner"); *Express*, BARCLAY, *supra* note 33 ("to represent in words, or by any of the imitative arts"); *Express*, DYCHE & PARDON, *supra* note 33 ("to speak or declare by word or writing"); *Express*, JOHNSON, *supra* note 33 ("[t]o represent by the imitative arts," "[t]o represent in words," "[t]o show or make known in any manner" "to exhibit by language"); *Express*, SHERIDAN, *supra* note 33 ("[t]o represent by any of the imitative arts," "to represent in words"); *Express*, WALKER, *supra* note 33 ("[t]o represent by any of the imitative arts," "to represent in words"); *Express*, WEBSTER, *supra* note 33 ("[t]o represent or show by imitation or the imitative arts," "[t]o show or make known").

42. *Thought*, WEBSTER, *supra* note 33; *see also Thought*, ASH, *supra* note 33 (providing a similar definition); *Thought*, BAILEY, *supra* note 33 (same); *Thought*, BARCLAY, *supra* note 33 (same); *Thought*, DYCHE & PARDON, *supra* note 33 (same); *Thought*, JOHNSON, *supra* note 33 (same); *Thought*, PERRY, *supra* note 33 (same); *Thought*, SHERIDAN, *supra* note 33 (same); *Thought*, WALKER, *supra* note 33 (same).

43. *Thought*, WEBSTER, *supra* note 33.

44. *Express*, JOHNSON, *supra* note 33.

45. *Speech*, 1 SAMUEL JOHNSON, A DICTIONARY OF THE ENGLISH LANGUAGE (London, W. Strahan et al., 4th ed. 1773) (emphasis added); *see also Thought*, WEBSTER, *supra* note 33 ("[T]o convey my *thoughts* to another person . . . I employ words that express my *thoughts*.").

From this distinction between thoughts and speech emerges the principle of speech certainty. Because speech is the thing uttered and expressed, speech can always be identified with certainty by the speaker at the moment it is spoken. Specifically, speech is that which the speaker produced as a result of transforming the inner workings of his mind into something he could "convey[] [from his] mind to another['s]."[46] It may be something spoken aloud, written down, sketched, drawn, painted, or performed.[47] But when it is uttered or expressed through "certain sounds, or written marks," the speaker will be able to point to what that speech was; he can identify it with certainty.[48]

Importantly, the fact that an audience may not be able to identify what the verbal speech *means* with any degree of confidence does not negate its speech certainty. In any communication or expression, something can always be lost in the translation from the speaker's thoughts to his speech, and from that speech to the audience's understanding of it.[49] But that translation loss applies only to the *message* underlying the speech, not the speech itself. Someone may, for example, misinterpret the statement "I saw her duck" as a reminiscence on seeing a friend's pet mallard, rather than a recollection of a boxer crouching to avoid a punch. But while the message's meaning may be misunderstood by others, the words themselves—"I," "saw," "her," and "duck," in that order—are imbued with indelible certainty. Those words—not their intended meaning or their misinterpretation—are the speaker's speech. And the speaker can identify what the speech is with total certainty, even if the audience may ultimately disagree on its meaning.

In summary, a comprehensive survey of Founding-era dictionaries reveals remarkably consistent definitions of speech. These definitions draw sharp distinctions between thoughts and speech, defining speech as the external manifestation of something that previously existed only in the speaker's mind. That external manifestation, by its very nature, will always be capable of certain identification by its speaker. Consequently, the Founding-era definitions of "speech," as understood by those who wrote it into the First Amendment and those who ratified it, reveal an inherent principle of speech certainty.

---

46. *Speech*, DYCHE & PARDON, *supra* note 33.

47. *See Express*, JOHNSON, *supra* note 33 ("by the imitative arts").

48. *Speech*, JOHNSON, *supra* note 45.

49. McCulloch v. Maryland, 17 U.S. (4 Wheat.) 316, 414 (1819) ("Such is the character of human language, that no word conveys to the mind, in all situations, one single definite idea . . . .").

B. History and the Original Public Meaning of "Speech"

At the Founding, there was no concept of speech uncharacterized by speech certainty. Speech was either spoken, written, or printed—forms of speech in which the speaker can be certain of what he said—or it wasn't speech. This is simply a historical fact given that in the Founding era "there were essentially three methods of communication: oral, unamplified speech; hand-written correspondence; and printed materials created using a printing press."[50] Each of these modes of communication are inherently and unavoidably characterized by speech certainty. By their very nature, they require that the speaker be able to identify with certainty what he said at the moment he said it. The act of speaking orally demands that something has been said aloud; writing demands something be written; printing that something be printed.

Consequently, the idea that "speech" might include that which a speaker couldn't know he said would be completely foreign to the Founders. Their analog era was necessarily limited to analog modes of communication. And those analog modes by their very nature abide by the principle of speech certainty.[51] Thus, while the original meaning of "the freedom of speech" protected by the First Amendment remains elusive, there can be no doubt that the "speech" it protected took the principle of speech certainty as a given.

This technological limitation on the meaning of "speech" is more than just a historical artifact—it was baked into the Founding generation's understanding of the First Amendment. Amid the Framers' heated disputes over the meaning of "the freedom of speech," two areas of rare consensus reveal that they were concerned about protecting "speech" characterized by speech certainty: the ban on prior restraints and permissibility of subsequent punishment. The history of these two bedrock bodies of speech law thus bolsters what the plain meaning of the text already exposes: The original public meaning of "speech" as used in the First Amendment incorporates the principle of speech certainty.

---

50. Ashutosh Bhagwat, *Posner, Blackstone, and Prior Restraints on Speech*, 2015 BYU L. REV. 1151, 1157 (2015).

51. Systems that created non-determinative speech existed in the Founding era and might be understood as producing uncertain speech. *See* James Grimmelmann, *There's No Such Thing as a Computer-Authored Work—And It's a Good Thing, Too*, 39 COLUM. J. L. & ARTS 403, 412 (2016) (describing a 1792 dice game for composing music). Given the nature of these systems, however, they can be considered early form of probabilistic traditional code, which, as explained in Part IV, is characterized by speech certainty. *See infra* Part IV.B.2. Thus, even the output of these machines and systems would, consistent with the principle of speech certainty, be protected under the First Amendment as the machine creator's speech.

1. Prior restraint

The Framers understood the First Amendment, at a minimum, to forbid prior restraints.[52] This blanket prohibition was written in direct contrast to British legal tradition and practice, in which publishers were historically subjected to licensing regimes imposing pre-publication review of any printed materials before their works could be distributed.[53]

The Licensing Act of 1662 is the paradigmatic case study of how prior restraint regimes operated in seventeenth-century England. Under the statute, a printer could only operate if it was granted a license from the royally chartered Stationer's Company.[54] Under that license, printers were required to submit all works to the Stationer's Company for pre-publication review to determine if they contained seditious or heretical material.[55] Once the Stationer's Company approved the work—or, more often, made irrebuttable deletions and edits to it—the Company would return the final draft to the printer for publication.[56] While the Framers debated over other parts of the First Amendment's meaning, they achieved a rare consensus that it barred such prior restraint.[57] A system of pre-publication review abridges "the freedom of speech."[58]

---

52. Michael I. Meyerson, *The Neglected History of the Prior Restraint Doctrine: Rediscovering the Link Between the First Amendment and the Separation of Powers*, 34 IND. L. REV. 295, 320-21 (2001) (noting the "widespread consensus" on "one critical principle," that "[l]iberty of the press must mean, at a bare minimum, no prior restraint"); *see, e.g.*, 2 THE DEBATES IN THE SEVERAL STATE CONVENTIONS ON THE ADOPTION OF THE FEDERAL CONSTITUTION AS RECOMMENDED BY THE GENERAL CONVENTION AT PHILADELPHIA IN 1787, at 449 (Jonathan Elliot ed., 2d ed. 1888) ("What is meant by liberty of press is, that there should be no antecedent restraint upon it.").

53. *See* Thomas I. Emerson, *The Doctrine of Prior Restraint*, 20 L. & CONTEMP. PROBS. 648, 650, 652 (1955).

54. *Id.* at 650.

55. *Id.* We distinguish licensing regimes—those which required publishers to submit potential works to a governmental authority for licensing prior to printing—and registration requirements in seventeenth century England. The Licensing Act of 1662 not only required that printed works be licensed prior to publication but also that any active printing presses be registered with the Stationers' Company. *See* The Licensing Act of 1662, 13 & 14 Car. 2, c. 33, § III (Eng.). In contrast to the history of licensing regimes, these registration requirements tell us little about the original public meaning of "speech" in the First Amendment. While licensing regimes illuminate when thoughts become "speech" by virtue of the pre-publication review process, registration regimes instead regulate when a person is at liberty to speak at all.

56. The Licensing Act of 1662, 13 & Car. 2, c. 33, § III (Eng.).

57. David S. Bogen, *The Origins of Freedom of Speech and Press*, 42 MD. L. REV. 429, 440-41 (1983); *see also* Near v. Minnesota, 283 U.S. 697, 713-14 (1931) (discussing the historical importance of the ban on prior restraints for the Framers).

58. We note that the Founding era's focus on pre-publication review is narrower than contemporary conceptions of it established in the early twentieth century. John Calvin

footnote continued on next page

This rejection of pre-publication review demonstrates how deeply the First Amendment is imbued with the principle of speech certainty. If prior restraint in the form of pre-publication review is among the First Amendment's primary targets, then the First Amendment's conception of "speech" presumes the existence of speech that is susceptible to such review. In the Founding era, the only such speech was that which a would-be speaker had already put into written or printed form—both of which are inherently characterized by speech certainty.[59] The writer knows what he wrote at the moment he wrote it; the same is true of the printer at printing. Indeed, the Stationer's Company could and did only review that which was already committed to paper, at which point the printer himself was necessarily certain of the contents of the printed speech.[60] Without that certainty, the pre-

---

Jeffries Jr., *Rethinking Prior Restraint*, 92 YALE L.J. 409, 414 (1983) (identifying *Near* as "the Supreme Court's first great encounter with prior restraint, and . . . the doctrine's leading precedent"). In *Near,* the Court refocused the question of whether a speech restriction qualified as a prior restraint by "[n]oting that the 'object and effect' of the statute was to 'suppress' *future* publication." Meyerson, *supra* note 52, at 337 (emphasis added) (quoting *Near*, 283 U.S. at 712). Since *Near*, the First Amendment's prohibition on prior restraint included not only pre-publication review of specific written or printed speech, but also preemptive injunctions on the publication of speech that were yet to be committed to paper. Meyerson, *supra* note 52, at 337. For speech certainty purposes, *Near* bifurcated prior restraint doctrine into two modes: (1) the prohibition on pre-publication review, which characterized Founding-era conceptions of prior restraint and concerns speech characterized by speech certainty, and (2) the prohibition of restraints on future speech, which was the subject of the *Near* opinion. *Near*, 283 U.S. at 711-13. This latter mode newly implicated First Amendment scrutiny even when there is no specific "speech" at issue. That is, the question it asks is not whether a speaker's "speech" falls under the First Amendment's protection, but whether a speaker has the right to speak at all. These "speech"-agnostic First Amendment protections thus provide a parallel form of First Amendment protection to traditional "speech"-specific speech protections (e.g., defamation, obscenity). They protect "the freedom of speech" independently from questions about whether a speaker's "speech" is protected (or even if it's a speaker's "speech" at all). The speech certainty principle is thus compatible with but is not relevant to these "speech"-agnostic protections (and vice versa). Its only concern is whether a speaker's purported speech is in fact "speech" under the First Amendment—a question *Near* left undisturbed. For more on the debates over the proper scope of the prior restraint doctrine, see Meyerson, note 52 above, at 338-42, and Jeffries, Jr., above, at 434-37.

59. *See* Bhagwat, *supra* note 50, at 1157 ("[T]here were essentially three methods of communication: oral, unamplified speech; hand-written correspondence; and printed materials created using a printing press.").

60. The Licensing Act of 1662, 13 & 14 Car. 2 c. 33, § III ("[N]o private person or persons whatsoever shall at any time hereafter print, or cause to be printed any book or pamphlet whatsoever, unless the same book and pamphlet, together with all and every the [sic] titles, epistles, prefaces, proems, preambles, introductions, tables, dedications, and other matters and things thereto annexed, be first entered in the book of the register of the company of stationers of London . . . ." (emphasis omitted)).

publication review process simply couldn't take place because there was no "speech" to review.

In short, the licensing regimes of seventeenth-century England depended for their very existence on the principle of speech certainty. And because the Framers intended to protect the "speech" enshrined by the First Amendment from such licensing regimes, their conception of that speech was inherently bounded by the principle of speech certainty.[61]

2. Subsequent punishment

Part and parcel of the prohibition against prior restraints, though, was the permissibility of subsequent punishment for at least some types of speech. While no government authority could ban purportedly criminal speech from the get-go, it could certainly punish it thereafter.[62] Blackstone's Commentaries illuminate the Founding-era consensus over the meaning of the term "speech." In Blackstone's authoritative conception, the freedom of the press "consist[ed] in laying no *previous* restraints upon publications, and not in freedom from censure for criminal matter when published."[63] Thus, once speech crossed the publication threshold "by the means of certain sounds, or written marks,"[64] punishments for it were fair game within hotly disputed limits.[65] That is, once a speaker has actually spoken, he may under certain circumstances be punished for the contents of his speech consistent with the First Amendment.

---

61. *See* Emerson, *supra* note 53, at 651-52. An important caveat to this discussion is that prior restraints were limited to written or printed speech—if only because prior restraints on oral speech were completely unfathomable to the Framers. *See* Bhagwat, *supra* note 50, at 1160-61 ("In no conceivable universe could the government require permission from censors before citizens could *speak* . . . ."). The concept is "difficult even to imagine in practice," writes the constitutional scholar Akhil Amar. AKHIL REED AMAR, THE BILL OF RIGHTS: CREATION AND RECONSTRUCTION 224 (1998). "Licensing the few printing presses that existed in the seventeenth and eighteenth centuries is one thing; but what would it *mean* to purport to license speakers and require official preclearance before one could open one's mouth?" *Id.* Indeed, "[t]he very idea of such a system is profoundly silly." Bhagwat, *supra* note 50, at 1160.

62. Bogen, *supra* note 57, at 440 n.52 (explaining why the First Amendment could not have been limited to prior restraint in the Founding era).

63. 4 WILLIAM BLACKSTONE, COMMENTARIES *151; *see also id.* at *152 ("A man (says a fine writer on this subject) may be allowed to keep poisons in his closet, but not publicly to vend them as cordials.").

64. *Speech,* JOHNSON, *supra* note 45.

65. The Alien and Sedition Act provided a constitutional flashpoint for the Framers in the years following the ratification of the First Amendment. *See* Sedition Act, ch. 74, 1 Stat. 596 (1798). Ultimately, however, the consensus view held that subsequent punishment for criminal speech was permissible under the speech clause. *See* Bogen, *supra* note 57, at 458 n.143.

The Framers' acceptance of subsequent punishment for speech reveals how deeply embedded the principle of speech certainty was in their conception of the First Amendment. Just as a regime of pre-publication review requires the existence of speech to review, a system of subsequent punishment requires the existence of speech to punish. In both cases, that speech must necessarily be characterized by speech certainty. Speech that wasn't characterized by speech certainty, e.g., unverbalized thoughts, simply couldn't be subjected to the necessary analysis.

Indeed, a brief survey of Founding-era defamation law—a common framework of subsequent punishment for speech—illustrates this dependence on speech certainty in practice. The Framers understood that, in light of its ban on prior restraint, libel (defamatory written or printed speech) and slander (defamatory oral speech) "could only be punished, but could not be prevented."[66] Yet to the Framers, defamation was a sufficiently undesirable category of speech that could be punishable by law without unconstitutionally "abridging the freedom of speech, or of the press."[67] Consequently, the cases enforcing defamation laws uniformly turned on the nature of the defamatory words at issue. The question was not whether the defendant said the defamatory words, but "whether the words be actionable."[68] Were they true?[69] Did they allege criminal conduct?[70] Were they spoken in private or to the public?[71] What the words meant, the context in which they were spoken,[72]

66. *See* Meyerson, *supra* note 52, at 309; *see also* AMAR, *supra* note 61, at 224.

67. U.S. CONST. amend. I. Other such categories included obscenity, fighting words, and incitement. *See* Brown v. Ent. Merchs. Ass'n, 564 U.S. 786, 791 (2011) ("From 1791 to the present, the First Amendment has permitted restrictions upon the content of speech in a few limited areas, and has never included a freedom to disregard these traditional limitations." (citation omitted)).

68. Shipp v. McCraw, 7 N.C. (3 Mur.) 463, 463 (1819).

69. People v. Croswell, 3 Johns. Cas. 337, 395 (N.Y. Sup. Ct. 1804) (Thompson, J., concurring) ("[I]f the truth of the facts charged as libellous [sic] should be made to appear, it would amount to a complete justification.").

70. *Shipp*, 7 N.C. (3 Mur.) at 466 (Henderson, J., concurring).

71. *See* Robert Post, *The Social Foundations of Defamation Law: Reputation and the Constitution*, 74 CALIF. L. REV. 691, 695 (1986) ("Defamation law should therefore not be concerned with purely private injuries which are independent of the market [for reputation].").

72. Even under the seventeenth and eighteenth century English legal regime, the punishments and remedies for defamation underline the speech certainty principle. When the Star Chamber was eventually abolished in 1641, British common-law courts—whose only permissible remedy was money damages—assumed jurisdiction over defamation suits. Meyerson, *supra* note 52, at 310. Accordingly, courts of equity—where injunctions could be issued—were denied jurisdiction over defamation cases. *Id.* The historical availability of remedies suggests that only litigants that suffered defamatory speech that was spoken with certainty were entitled to a remedy, but would-be speakers would not and could not be enjoined preemptively.

and the extent of their harm were the consistent focus of the courts.[73] That the defamatory speech was characterized by speech certainty—that is, that the defendant said what he said and knew he said it when he did—was a threshold matter in such cases. Were it not, there would be no case for a court to try.

At its most basic level then, the Founders understood that to punish speech, one must know what that speech is. In the Founding era, that meant that it must be imbued with speech certainty.

C. Speech Certainty and the Purpose(s) of the First Amendment

1. Speech certainty and the leading theories of the First Amendment

The principle of speech certainty is not only embedded within the Speech Clause's text and history but supported by the First Amendment's core purposes. Because a consensus definition of "the freedom of speech" is out of reach,[74] understanding *why* the Constitution protects this freedom has offered another means of determining its proper scope.[75] Scholars thus tend to invoke one of three motivating theories of the First Amendment's protection of the freedom of speech: the autonomy theory, the democratic-process approach, and the marketplace rationale.[76]

All of these purpose-based theories work—sometimes individually, but oftentimes in tandem—to justify different boundaries of protected and unprotected speech under the First Amendment.[77] The precise meaning of these purposes and their implications for how courts should interpret the First Amendment have been, and remain, the subject of much debate.[78] At bottom, though, they are best understood as debates over which ideals ought to underpin the interpretation of the First Amendment. On that score, then, we agree with Tim Wu's conclusion that relying on these theories to "distinguish[]

73. *See, e.g.*, RODNEY A. SMOLLA, 1 LAW OF DEFAMATION § 4:1 (2d ed. 2024) ("In the final analysis what matters most is that the statement upon which the claim of defamation is based be a statement that would truly matter to some segment of the community.").

74. *See supra* note 27 (collecting sources).

75. David S. Han, *The Value of First Amendment Theory*, 2015 U. ILL. L. REV. SLIP OP. 87, 88-89 (2015), https://perma.cc/Q7YY-X3WH ("[T]he drive to identify a unifying theory of the First Amendment is ultimately rooted in a search for determinacy and coherence . . . .").

76. *See* David S. Han, *Autobiographical Lies and the First Amendment's Protection of Self-Defining Speech*, 87 N.Y.U. L. REV. 70, 89-90 (2012).

77. *See* Tim Wu, *Machine Speech*, 161 U. PA. L. REV. 1495, 1507 (2013).

78. *See, e.g.*, Dale Carpenter, *Editor's Note*, 27 CONST. COMMENT. 249 (2011) (preceding two essays that "defend their separate visions of autonomy as the basis of American free speech protection").

a subset of protected speech from all communications is necessarily a normative project."[79]

The extent to which normative values offer a practical guide for our interpretation of the First Amendment is beyond the scope of this paper.[80] But accepting that they do, at least to some extent, we provide in this Subpart a high-level overview of (in the authors' view) the most salient of the different theories: the autonomy justification.[81] In doing so, we show that the

79. Wu, *Machine Speech*, *supra* note 77, at 1506.

80. *See, e.g.*, William Baude & Michael Stokes Paulsen, *The Sweep and Force of Section Three*, 172 U. PA. L. REV. 605, 613 (2024) ("[I]t is (or should be) basic constitutional law that it is the enduring text of the Constitution that supplies the governing rule, not the ostensible 'purpose' or specific historical situation for which the text was written." (emphasis omitted)); C. Edwin Baker, *Autonomy and Free Speech*, 27 CONST. COMMENT. 251, 269 n.27 (2011) ("[L]awyers and judges normally and probably wisely avoid explicit engagement with moral philosophy. That fact does not deny, however, that they necessarily rely on normative premises . . . .").

81. The other two theories also arguably support the recognition of the speech certainty principle. Some scholars, for example, have suggested that any conception of the marketplace of ideas as a justification for the First Amendment is limited to a marketplace of ideas propounded by people. Jared Schroeder, *Marketplace Theory in the Age of AI Communicators*, 17 FIRST AMEND. L. REV. 22, 22-23 (identifying "the nature of human actors who take part in communicating ideas" as a core assumption of the marketplace theory); Morgan N. Weiland, *First Amendment Metaphors: The Death of the "Marketplace of Ideas" and the Rise of the Post-Truth "Free Flow of Information"*, 33 YALE J.L. & HUMANS. 366, 397 (2022) ("[T]he marketplace model . . . imagined the press and journalists as rational actors and impliedly human."); *see also* Whitney v. California, 274 U.S. 357, 375 (1927) (Brandeis, J., concurring) ("[T]he final end of the State [is] to make men free to develop their faculties."); Rodney A. Smolla, *The Meaning of the "Marketplace of Ideas" in First Amendment Law*, 24 COMM. L. & POL'Y, 437, 465 (2019) (identifying *Whitney* as one of "[t]he two seminal articulations of the marketplace theory"). Because, as this Article explains, every form of expression recognized as protected speech is necessarily characterized by speech certainty, the marketplace theory of the First Amendment can be understood to limit its protection to such speech.

The democratic process approach is similarly supportive of the speech certainty principle. Under this theory, the First Amendment protects "the freedom of those activities of thought and communication by which we 'govern.'" Alexander Meiklejohn, *The First Amendment is an Absolute*, 1961 SUP. CT. REV. 245, 255 (1961). In other words, it protects political speech. Tsesis, *supra* note 22, at 1035-36 (describing how the theory distinguishes between "political and private speech"). According to Alexander Meiklejohn, the leading proponent of the democratic-process theory, speech is political if it reflects citizens' actual perspectives on matters of self-government; only then does it earn the protection of the First Amendment. *See* ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT 27 (1948) ("If [such perspectives] are responsibly entertained by anyone, we, the voters, need to hear them."). Uncertain speech—"speech" that a speaker does not know he said—cannot be said to reflect a political perspective "responsibly entertained by anyone." *Id.* Thus, if the First Amendment protects speech that *is not* characterized by speech certainty, it would inevitably protect something other than the political speech of its citizens—speech that, under this theory, the First Amendment does not protect.

distinctions that the theory draws between protected and unprotected speech support, if not demand, recognition of the speech certainty principle.

Under autonomy theory, "use of communication as a means of self-expression, self-realization, and self-fulfillment" is "the principal function of the First Amendment."[82] Baker posits that the purpose of "the freedom of speech" is to safeguard the individual's autonomy over what she chooses to say and not to say, so long as those choices don't impinge on anyone else's expressive autonomy.[83] Only if everyone is free to express themselves can the Constitution's primary project of guaranteeing individual and collective self-determination succeed.[84]

If autonomy theory dictates that the First Amendment's purpose is to protect the speaker's self-expression, then it compels recognition of the principle of speech certainty. That is, for the idea of expressive autonomy to have any meaning, a speaker's choices about what to express (and not to express) must be his own.[85] And if that expression is truly his own, he must know what it is with certainty when he expresses it. If a speaker can't identify his own expression at the moment he expresses it, how can he even know whether he chose to express himself at all? And, in the event that he did, how can he know whether that expression corresponded to his expressive choices? Without speech certainty, the purported "speech" simply cannot be what the Supreme Court has called a "manifestation of individual freedom or choice."[86] Speech without speech certainty severs the link between a speaker and his self-expression that the autonomy theory demands.

The autonomy theory's limitation of "the freedom of speech" to acts of self-expressive liberty therefore *depends* on the speech certainty principle. To protect speech that isn't characterized by speech certainty—speech a speaker

82. *See* Baker, *Autonomy and Free Speech*, *supra* note 80, at 274 (quoting First Nat'l Bank v. Bellotti, 435 U.S. 765, 804 (1978) (White, J., dissenting)). Indeed, Baker identifies "[t]he poster child of autonomy theory" as "the Court's opinion in *Barnette*." *Id.* at 270. In *West Virginia State Board of Education v. Barnette*, the Court "gave a ringing endorsement of the school children's right to abstain from saluting the flag on the basis of First Amendment protected liberty." *Id.* As Baker summarizes, the majority held that "[c]ompulsion directly abridged children's liberty" and focused the analysis on the children's "expressive autonomy" rather than the impact of the speech on the listener (the focus of the marketplace rationale) or political discourse (the focus of the democratic-process approach). *Id.* at 271.

83. *See id.* at 254.

84. *See id.* at 265.

85. Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 573 (1995) (per curiam) ("[T]he fundamental rule of protection under the First Amendment [is] that a speaker has the autonomy to choose the content of his own message.").

86. Baker, *Autonomy and Free Speech*, *supra* note 80, at 274 (quoting *Bellotti*, 435 U.S. at 805 (White, J., dissenting)).

can't identify with certainty at the moment he says it—would be to protect something other than an act of self-expressive liberty. Under the autonomy theory, then, such protection would improperly cover "speech" that is beyond the scope of the First Amendment.

2. Speech certainty and the purpose of the First Amendment's categorical exceptions

As Tim Wu has observed, the various theories of the First Amendment only illuminate the boundaries between speech and non-speech so much. They provide judges with direction, but not directions for judges.[87] On that score, the "categorical approach" lights a clearer path.[88] By expressly defining categories of speech that are unprotected—e.g., obscenity, true threats, defamation—this approach tells us "where the First Amendment is 'on' and where it is 'off.'"[89] As a result, it allows us another angle from which to explore the First Amendment and whether it supports the adoption of the principle of speech certainty.

While some dispute the virtues of the categorical approach,[90] the Court has in recent decades entrenched its supremacy as a doctrinal guide for content-based restrictions on speech.[91] These restrictions, known as the categorical exceptions to the First Amendment, presuppose the principle of speech certainty. Indeed, the way the Court describes them makes no sense without it.

---

87. Wu, *Machine Speech*, *supra* note 77, at 1508 ("The line between communications and speech exists as much as a judicial necessity as anything else. But how and where is that line drawn?").

88. *Id.* at 1509.

89. *Id.* Alongside these "formal" exclusions, Wu also describes the "informal" exclusion of non-expressive conduct from First Amendment protection. *Id.* at 1509-10. We address the relationship between speech certainty and the expressive conduct doctrine in Part II.B. Wu also describes medium-specific "inclusion categories," such as the evolving First Amendment protection for films, as part of the categorical approach. *Id.* at 1512. "An inclusion works this way: When the medium is used, the communicator is presumptively a speaker." *Id.* at 1511. Under the principle of speech certainty, so long as a speaker knows what he says when he says it, his speech is "speech" within the meaning of the First Amendment, regardless of the medium through which it is communicated. *Cf. id.* (recognizing the "presumption" of speech when communicated via certain media). Whether that speech is protected or not, however, is subject to the Court's "inclusion categories." *See id.*

90. Wu, *Machine Speech*, *supra* note 77, at 1509; *cf.* Gregory P. Magarian, *The Marrow of Tradition: The Roberts Court and Categorical First Amendment Speech Exclusions*, 56 WM. & MARY L. REV. 1339, 1341-45 (2015) (critiquing the Roberts Court's development of the categorical approach doctrine).

91. United States v. Alvarez, 567 U.S. 709, 717 (2012); United States v. Stevens, 559 U.S. 460, 468-72 (2010).

In recent years, the Court has justified the categorical exceptions by deferring to its "historical foundation in the Court's free speech tradition."[92] This free speech tradition tolerates certain content-based restrictions on speech that, as Justice Roberts articulated in *United States v. Stevens*, share a dubious distinction:

> [W]ithin these categories of unprotected speech, "the evil to be restricted so overwhelmingly outweighs the expressive interests, if any, at stake, that no process of case-by-case adjudication is required," because "the balance of competing interests is clearly struck."[93]

Although the Court rejected the use of "an ad hoc balancing of relative social costs and benefits" to determine whether *other* content-based restrictions on speech may be constitutional, it explained that the "historic and traditional" categories of speech excluded from the First Amendment's protection strike a uniquely "overwhelming[]" imbalance that justifies their exclusion.[94] Indeed, when discussing the categorical exceptions, the Court has repeatedly emphasized that "any benefit that may be derived from them is clearly outweighed by the social interest in order and morality."[95]

In other words, like the vast majority of Founding-era theories of punishment, the categorical exceptions—and specifically the permissible punishment for their utterance—are, at least in part, justified by deterrence.[96]

---

92. *Alvarez*, 567 U.S. at 718.

93. *Stevens*, 559 U.S. at 470 (quoting New York v. Ferber, 458 U.S. 747, 763-64 (1982)).

94. *Id.* at 468, 470; *Ferber*, 458 U.S. at 763-64 ("[I]t is not rare that a content-based classification of speech has been accepted *because* it may be appropriately generalized that within the confines of the given classification, the evil to be restricted so overwhelmingly outweighs the expressive interests, if any, at stake, that no process of case-by-case adjudication is required." (emphasis added)). In *Stevens*, Justice Roberts refutes the idea that categorical exceptions are justified by "this 'balance of competing interests' *alone*." 559 U.S. at 471 (emphasis added) (quoting *Ferber*, 458 U.S. at 763-64). To support this argument, he explains that the recognition of child pornography as a categorical exception was also justified by the historic lack of protection for speech integral to criminal conduct. *Id.* at 471-72. Thus, while each categorical exception may have multiple justifications for its exclusion from First Amendment protection, Justice Roberts acknowledges that the common thread across all of them is that "any benefit that may be derived from them is clearly outweighed by the social interest in order and morality." *Id.* at 470 (quoting R.A.V. v. St. Paul, 505 U.S. 377, 383 (1992)).

95. *Stevens*, 559 U.S. at 470. ("[T]his Court has often *described* historically unprotected categories of speech as being 'of such slight social value as a step to truth that any benefit that may be derived from them is clearly outweighed by the social interest in order and morality.'" (quoting *R.A.V.*, 505 U.S. at 383)). For a deeper discussion on this topic, see Magarian, *supra* note 90, at 1346-48.

96. Dawinder S. Sidhu, *Towards the Second Founding of Federal Sentencing*, 77 MD. L. REV. 485, 489 (2018) ("At the country's inception, sentencing was driven by principles of deterrence, with retribution playing a role in only the most heinous of crimes." (citing DAVID J. ROTHMAN, PERFECTING THE PRISON: UNITED STATES, 1789-1865, *in* THE

footnote continued on next page

For example, the goal of refusing to protect defamation is to minimize the frequency with which people defame others.[97] If Harry knows he can be punished for defaming Sally, he will be less likely to circulate flyers that falsely allege she's a crook. In the context of protected speech, this logic is called the unconstitutional chilling of speech;[98] but for the categorical exceptions, chilling unprotected speech is the intended and desired effect to promote "the social interest in order and morality."[99] As a result, courts have deemed these categories of speech so "overwhelmingly" contrary to the broader purposes of the freedom of speech that we carve them out from its protection.[100]

This deterrence-based justification for the categorical exceptions thus presupposes the speech certainty principle. Only if speakers know what they say when they say it—and, as a result, can choose to say one thing instead of another—does it make sense to justify the punishment for certain types of speech on the basis that it makes people less likely to utter that speech. This rationale presumes that the threat of financial damages (or worse) will steer Harry away from circulating those flyers about Sally. But if Harry (for some reason) can't know what the flyers he has written will say when he circulates them, he might defame Sally—and only discover that he has done so at some later time.[101] In that scenario, the threat of punishment wouldn't deter Harry because deterrence has no influence over a speaker who doesn't know what he says when he says it. Without speech certainty, the "evil to be restricted" can't be discouraged; the primary historic and traditional aim of the "historic and traditional" categorical exceptions would be gutted.[102] Instead, the categorical exception framework necessarily assumes a speech certainty principle.

* * *

In summary, a review of the text, history, and underlying purposes of the First Amendment compels the recognition that Founding-era conceptions of

---

OXFORD HISTORY OF THE PRISON: THE PRACTICE OF PUNISHMENT IN WESTERN SOCIETY 111, 111-13 (Norval Morris & David J. Rothman eds., 1995))).

97. *See, e.g.*, Rosenblatt v. Baer, 383 U.S. 75, 86 (1966) ("Society has a pervasive and strong interest in *preventing* and redressing attacks upon reputation." (emphasis added)).

98. *See, e.g.*, Marcone v. Penthouse Int'l Mag. for Men, 754 F.2d 1072, 1080-81 (3d Cir. 1985) ("The Court granted First Amendment protection to negligently false statements in order to afford the media the 'breathing space' necessary to avoid a chilling effect on constitutionally valuable speech . . . ." (quoting New York Times Co. v. Sullivan, 376 U.S. 254, 271-72 (1964))).

99. *Stevens*, 559 U.S. at 470 (quoting *R.A.V.*, 505 U.S. at 383).

100. New York v. Ferber, 458 U.S. 747, 763-64 (1982).

101. *See infra* note 140 (addressing how the principle of speech certainty applies to *Smith v. California*, 361 U.S. 147 (1959)).

102. *Stevens*, 559 U.S. at 468, 470 (first quoting *Ferber*, 458 U.S. at 763; and then quoting Simon & Schuster, Inc. v. Members of the N.Y. State Crime Victims Bd., 502 U.S. 105, 127 (1991) (Kennedy, J., concurring in the judgment)).

"speech" were bounded by the speech certainty principle. First, Founding-era dictionaries unequivocally reveal that the textual meaning of "speech" required that a speaker know with certainty what he said when he said it for it to qualify as "speech." Indeed, the original public meaning of the term presupposed that the speaker manifest his thoughts through the spoken, written, or printed word before they crystallized into what the Founders would have considered "speech" within the meaning of the First Amendment.

Second, historical consensus of what, at a minimum, "the freedom of speech" prohibited (prior restraint) and what it allowed (limited forms of subsequent punishment) likewise presuppose the principle.[103] Without speech committed to the page—necessarily characterized by speech certainty—"the freedom of speech" couldn't be abridged by pre-publication review. Nor could any speech at the time of the Founding be punished after-the-fact if a speaker had not previously spoken with the requisite certainty. Any other conception of speech would have been unfathomable to those in the Founding era, limited as they were to oral, written, and printed expression.

Finally, setting aside Founding-era conceptions of speech, the First Amendment's purposes confirm that the principle of speech certainty is more than just a product of the Framers' time. And the Court's consistent justification of the categorical exceptions as, in part, fostering "the social interest in order and morality" assumes that the unprotected speech within its purview is characterized by speech certainty; if speakers can't know what they say when they say it, they can't be deterred from saying it.[104]

Thus, three distinct modes of constitutional interpretation—text, history, and purpose—all counsel toward recognizing the as-yet unrecognized principle of speech certainty as an inherent limitation to the scope of the First Amendment. In the next Part, we shift from text, history, and purpose to Supreme Court precedent to show that modern expansions of the First Amendment compel the same conclusion.

II. Speech Certainty and First Amendment Jurisprudence

Since the Founding, both the development of new technologies and an evolving understanding of the purpose of the First Amendment have motivated the Supreme Court to expand the doctrine. Of particular importance are two branches of speech law: editorial discretion and expressive

103. For discussion on how contemporary First Amendment jurisprudence has expanded the scope of "the freedom of speech" to do more than protect "speech," see note 58 above.

104. *Stevens*, 559 U.S. at 470 (quoting *R.A.V.*, 505 U.S. at 383).

conduct.[105] In this Part, we show that the baseline requirements of both doctrines reflect the same presumption of speech certainty as all other forms of protected speech: To earn First Amendment protection, the speaker must know the contents of her speech at the moment she says it, or in the case of editorial discretion, the moment the speech is published. Here, we dig into the Court's editorial discretion and expressive conduct case law to illustrate that even as the doctrine expanded the scope of the First Amendment's protection, it kept the principle of speech certainty intact.[106]

A. Editorial Discretion

Over the last century, the doctrine of editorial discretion has blossomed from an implicit principle into an essential right that "all recognize as fundamentally protected."[107] Justice Hugo Black planted its first seeds in a footnote of *Associated Press v. United States,* a 1945 case acknowledging that a newspaper wiring service has the right not to publish "anything which their

105. We acknowledge that, given the unique nature of traditional computer source code, some lower courts have analyzed computer code itself—as distinguished from the output of algorithmic models—as pure speech. That analysis—the expressive-functional analysis—asks courts to analyze the general expressiveness and functionality of the source code in question and determine whether the expressive parts outweigh the functional ones such that the First Amendment attaches. If source code "combin[es] nonspeech and speech elements, i.e., functional and expressive elements," it is still entitled to protection. Universal City Studios, Inc. v. Corley, 273 F.3d 429, 448 n.21, 451 (2d Cir. 2001) (noting that because computer code is a form of language and the preferred method of communication among computer programmers, courts have held that computer code is pure speech); *see also* Junger v. Daley, 209 F.3d 481, 485 (6th Cir. 2000) ("[C]omputer source code is an expressive means for the exchange of information and ideas about computer programming . . . ."); Bernstein v. U.S. Dep't of Just., 176 F.3d 1132, 1142 (9th Cir. 1999) ("[S]ource code is utilized by those in the cryptography field as a means of expression . . . ."). *See generally* Xiangnong Wang, *De-Coding Free Speech: A First Amendment Theory for the Digital Age*, 2021 WIS. L. REV. 1373 (2021) (discussing the history and implications of the code-as-speech doctrine). The "code is speech" line of cases is beyond the scope of this paper; instead, we ask whether the *output* of machine learning algorithms is speech under the First Amendment. For speech certainty purposes, however, we note that whatever protection is afforded to it, the process of writing code is functionally identical to the writing of other speech; by writing it, the programmer can be certain of the code they have written.

106. We do not argue that the Court did so deliberately or consciously. Instead, just as the Framers' notions of speech in the eighteenth century were limited to that which was characterized by speech certainty, so too was the Court's in the twentieth. Upon scrutiny, their analysis therefore reveals the same implicit presumption that First Amendment Speech was characterized by speech certainty, and the First Amendment—including the doctrines of editorial discretion and expressive conduct—protected speech.

107. Denver Area Educ. Telecomms. Consortium, Inc. v. FCC, 518 U.S. 727, 822 (1996) (Thomas, J., concurring in part and dissenting in part).

'reason' tells them should not be published."[108] In a series of cases establishing the First Amendment rights of newspapers and broadcasters in the 1970s, the Supreme Court referred to that right as "journalistic discretion," "editorial judgment," and "editorial discretion."[109] By the 1990s, it extended the right to cable television operators and, finally, beyond traditional media companies.[110] And most recently, in *Moody v. NetChoice, LLC*, the Court indicated that—pending further factual development— the doctrine might apply to social-media platforms in the digital age.[111]

Despite this rich history, the specifics of the doctrine's application remain somewhat hazy.[112] There can be no doubt that the First Amendment forbids the government from compelling speakers who have the right of editorial discretion from publishing "that which 'reason' tells them should not be published."[113] But what is it about a speaker's speech that makes her eligible for that right? Until recently, the question had never been directly answered in any of the Supreme Court's editorial discretion cases individually.[114] In *Moody*, however, the Court summarized what a review of the cases makes clear.[115]

---

108. Associated Press v. United States, 326 U.S. 1, 20 n.18 (1945); *see also* Mia. Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 256 (1974) (noting that the origins of editorial discretion began with *Associated Press*).

109. Evelyn Douek & Genevieve Lakier, *Rereading "Editorial Discretion,"* KNIGHT FIRST AMEND. INST. (Oct. 24, 2022), https://perma.cc/SR8N-TVEK.

110. Turner Broad. Sys. v. FCC, 512 U.S. 622, 636 (1994) (cable operators); Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 569-70 (1995) (parade organizers).

111. 144 S. Ct. 2383, 2405 (2024) (explaining that the doctrine of editorial discretion grants protection "[f]or a paper, and for a platform too"). *But see id.* at 2403 ("These cases, to be sure, are at an early stage; the record is incomplete even as to the major social-media platforms' main feeds, much less the other applications that must now be considered."); *id.* at 2422 (Alito, J., concurring in the judgment) (arguing that the Court's discussion of editorial discretion and its application to social-media platforms is "nonbinding dicta").

112. Douek & Lakier, *supra* note 109 ("It is true that the Court has never been particularly clear about how to define 'editorial discretion.'"); *see also* Stuart Minor Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445, 1491-92 (2013).

113. *Tornillo*, 418 U.S. at 256 ("Compelling editors or publishers to publish that which reason tells them should not be published is what is at issue in this case." (internal quotations omitted)); *see also Moody*, 144 S. Ct. at 2402 ("An entity 'exercis[ing] editorial discretion in the selection and presentation' of content is 'engage[d] in speech activity.' . . . Deciding on the third-party speech that will be included in or excluded from a compilation—and then organizing and presenting the included items—is expressive activity of its own." (quoting Ark. Educ. Television Comm'n v. Forbes, 523 U.S. 666, 674 (1998))).

114. *See* Ashutosh Bhagwat, *Do Platforms Have Editorial Rights?*, 1 J. FREE SPEECH L. 97, 100-01 (2021) ("Editorial rights are of course well established with respect to traditional print newspapers. But who else enjoys such rights is unclear.").

115. *See Moody*, 144 S. Ct. at 2400-06 (summarizing the editorial discretion cases).

Collectively, the editorial discretion cases reveal at least the following common thread, consistently woven throughout them over half a century: A speaker is eligible for the First Amendment right of editorial discretion when she creates a compilation of speech,[116] exercises her judgment about what speech to include or exclude in that compilation, and publishes the compilation.[117]

These building blocks of editorial discretion fundamentally depend on the speech certainty principle. First, the Court consistently emphasizes the "right as a private speaker to shape its expression by speaking on one subject while remaining silent on another."[118] Such a right only has meaning if a speaker knows the content of her speech and can determine whether she has chosen to speak on a subject or remain silent on it. Second, the Court uniformly imposes a "publication" requirement on speech seeking protection under the doctrine of editorial discretion.[119] This requirement means that the editorial process must reach a final editorial judgment—a "publication" in some form—prior to

116. *Id.* at 2401. Because there is broad consensus about this prerequisite, we do not flesh it out in the body of the Article as we do for the other two requirements. But to eliminate any doubt: As the Court explained in *Hurley*, "under our precedent," the First Amendment protects any "private speaker" who "combin[es] multifarious voices" in a "communication," whether the speaker "generate[s], as an original matter, each item featured in the communication" or not. Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 569-70 (1995); *see also* Turner Broad. Sys., Inc. v. FCC., 512 U.S. 622, 675 (1994) (O'Connor, J., concurring in part and dissenting in part) ("Selecting which speech to retransmit is . . . no less communication than is creating the speech in the first place."). The "precedent" invoked in *Hurley* is the editorial discretion line of cases. 515 U.S. at 570 (citing *Turner*, 512 U.S. at 636; and *Tornillo*, 418 U.S. at 258). These cases concern the right of newspapers to choose what combination of articles, opinions, and advertisements to publish in its pages, and of broadcasters and cable television operators to decide what determines the combination of programming to feature in their transmissions. *See generally Tornillo*, 418 U.S. 241 (articles and opinions); Pittsburgh Press Co. v. Pittsburgh Comm'n on Hum. Rels., 413 U.S. 376 (1973) (advertisements); Columbia Broad. Sys., Inc. v. Democratic Nat'l Comm., 412 U.S. 94 (1973) (broadcast spot announcements); *Turner*, 512 U.S. at 622 (local broadcasting). Thus, it is compilations of speech that are protected by the First Amendment through the doctrine of editorial discretion. Consequently, the first step in the analysis of determining whether the doctrine may be invoked is to determine whether such a compilation exists.

117. *Moody*, 144 S. Ct. at 2402 (explaining that editorial discretion protects "[d]eciding on the . . . speech that will be included in or excluded from a compilation—and then organizing and presenting the included items").

118. *Hurley*, 515 U.S. at 574; *see also Moody*, 144 S. Ct. at 2401 (describing *Hurley* as the "capstone" of the editorial discretion cases).

119. *Moody*, 144 S. Ct. at 2400 (explaining that editorial discretion protects "presenting a curated compilation of speech"); *see id.* at 2402 (same); Herbert v. Lando, 441 U.S. 153, 178 (1979) (Powell, J., concurring) ("[W]hatever protection the exercise of editorial judgment enjoys depends entirely on the protection the First Amendment accords the product of this judgment, namely, published speech." (internal quotations omitted)).

earning First Amendment protection. In other words, it identifies *when* the speaker must know the contents of her speech as the time of publication. Together, these two elements of the Court's analysis show that the doctrine of editorial discretion adheres to the principle of speech certainty.

1. Exercises judgment about the contents of the compilation

When Justice Black planted the seeds of editorial discretion in *Associated Press*, it was not a blanket speech protection for the creators of compilations, but a prohibition forbidding the government from compelling them to publish "anything which their 'reason' tells them should not be published."[120] In *Tornillo*, the "seminal" editorial discretion case,[121] the Court reaffirmed this conception of the doctrine.[122] By emphasizing the editor's "reason," the Court centers the analysis on the editor's judgment about the contents of the compilation—the right to ensure that her speech reflects her decisions about what it ought to include.[123] The editor decides what gets published in the newspaper and what gets left on the cutting room floor. Inherent to this right, then, is the editor's certainty about the contents of her speech. If the editor's speech—the product of her editorial discretion—weren't characterized by speech certainty, she couldn't know whether it actually reflects her judgment. For nearly eighty years, this presumption of the editor's control over the contents of the speech—and thus her certainty as to it—has been a fixture of the Court's analysis in editorial discretion cases.[124]

In *Columbia Broadcasting System, Inc. v. Democratic National Committee*, the Court's first editorial discretion case following *Associated Press*, the Court considered whether a broadcaster could be compelled to accept an anti-war advertisement.[125] Finding it could not, the Court forcefully defended the

120. Associated Press v. United States, 326 U.S. 1, 20 n.18 (1945).

121. *Moody*, 144 S. Ct. at 2400.

122. *See Tornillo*, 418 U.S. at 256, 258 ("Compelling editors or publishers to publish that which "'reason" tells them should not be published' is what is at issue in this case." (quoting Grosjean v. Am. Press Co., 297 U.S. at 233, 244-45)).

123. *Id.* at 256; *see also Moody*, 144 S. Ct. at 2400 ("Forcing the paper to print what 'it would not otherwise print . . . intru[ded] into the function of editors[,]' . . . [f]or that function was, first and foremost, to make decisions about the 'content of the paper' and '[t]he choice of material to go into' it." (quoting *Tornillo*, 418 U.S. at 256, 258)); *see Tornillo*, 418 U.S. at 261 (White, J., concurring) (noting that editors' "decision[s] as to what copy will or will not be included in any given edition" are "the very nerve center of a newspaper").

124. *See Tornillo*, 418 U.S. at 256 (tracing editorial discretion from its origins in *Associated Press*); *Moody*, 144 S. Ct. at 2400-03 (tracing editorial discretion from *Tornillo* to the present).

125. 412 U.S. 94, 97-98 (1973).

"editorial judgment" and "journalistic discretion" of the "editors and publishers" of both newspapers and broadcasters.[126] These editors exercise their judgment and discretion through the act of editing, which the Court defined in plain and ordinary terms as the "selection and choice of material."[127] When the broadcaster chose not to accept an anti-war advertisement, it had selected and chosen the material to go into their broadcasts, and chosen to exclude the advertisement. Their decision was "expressly based on a journalistic judgment."[128] As Justice Black articulated in *Associated Press*, the editors were persuaded by their reason not to publish such advertisements, so they refrained from doing so.[129] The decision of the editors in *Columbia Broadcasting System* to exclude certain such ads was therefore a protected exercise of editorial discretion under the First Amendment.[130]

The following term, the Court decided *Miami Herald Publishing Company v. Tornillo*, in which a newspaper had refused to publish a response from a political candidate in violation of a Florida statute.[131] In striking down the law, the Court observed that the law "br[ought] about a confrontation with the express provisions of the First Amendment and the judicial gloss on that Amendment developed over the years."[132] This "gloss" referred to the doctrine of editorial discretion. Tracing its history "beginning with *Associated Press*" through *Columbia Broadcasting System*, the Court articulated an "expressed sensitivity as to . . . the compulsion exerted by government on a newspaper to print that which it would not otherwise print."[133] Summarizing that sensitivity, the *Tornillo* decision set forth the clearest definition of the doctrine of editorial discretion in the Court's history:

> The choice of material to go into a newspaper, and the decisions made as to limitations on the size and content of the paper, and treatment of public issues

126. *Id.* at 111, 117-18.

127. *Id.* at 124.

128. *Id.* at 118 ("[T]hat 10- to 60-second spot announcements are ill-suited to intelligible and intelligent treatment of public issues.").

129. *See* Associated Press v. United States, 326 U.S. 1, 20 n.18.

130. *Columbia Broad. Sys.*, 412 U.S. at 118, 121. The Court did not, however, recognize an absolute right at this stage of doctrinal development. *See id.* at 119 (refusing to definitively answer whether the First Amendment precludes the government from influencing editorial policies).

131. 418 U.S. 241, 341-42 (1974).

132. *Id.* at 254.

133. *Id.* at 256.

> and public officials—whether fair or unfair—constitute the exercise of editorial control and judgment.[134]

Once again, the effect of the editors' decision was central to the analysis; their "choice of material" did not include the candidate's reply and, as a result of that editorial judgment, they could be certain that the next day's edition of the newspaper would not include it.

Over the ensuing decades, the Court reinforced these same principles as it recognized that the right to editorial discretion also applied to cable operators,[135] which was beyond the traditional media to any editor of a compilation.[136] Over half a century, the Court handed down a line of Supreme Court cases applying the doctrine of editorial discretion in different contexts: newspapers, broadcasters, cable programmers and operators, and, finally, parade organizers.[137]

In *Moody*, the Court tentatively uncorked the doctrine for the first time in over a quarter century to frame its discussion about how it might apply to social-media platforms.[138] It summarized the lessons from the earlier cases, stating flatly that:

> Deciding on the . . . speech that will be included in or excluded from a compilation—and then organizing and presenting the included items—is expressive activity of its own. And that activity results in a distinctive expressive product. When the government interferes with such editorial choices—say, by ordering the excluded to be included—it . . . overrid[es] a party's expressive choices . . . [and] confronts the First Amendment.[139]

This concise summary in *Moody v. NetChoice* reflects the central premise of each and every editorial discretion case: Editors must know with certainty that

134. *Id.* at 258; *see also id.* at 261 (White, J., concurring) (noting editor's "decision[s] as to what copy will or will not be included in any given edition" are "the very nerve center of a newspaper").

135. FCC v. Midwest Video Corp., 440 U.S. 689, 707 (1979) (noting that cable operators "share with broadcasters a significant amount of editorial discretion regarding what their programming will include").

136. *See* Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 575 (1995) ("[W]hatever the reason, it boils down to the choice of a speaker not to propound a particular point of view, and that choice is presumed to lie beyond the government's power to control.").

137. *See Tornillo*, 418 U.S. 241 (newspapers); Columbia Broad. Sys., Inc. v. Democratic Nat'l Comm., 412 U.S. 94 (broadcasters); Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622 (1994) (cable programmers and operators); *Hurley*, 515 U.S. 557 (parade organizers).

138. Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2402 (2024). While the Court did invoke editorial discretion in *Manhattan Community Access Corporation v. Halleck*, the case does not address the meaning or scope of the doctrine. 587 U.S. 802, 805, 812-16 (2019).

139. *Moody*, 144 S. Ct. at 2402.

their speech includes what they intended to include and excludes what they intended to exclude.[140]

2. Publishes the compilation

The editorial discretion cases not only require that editors know with certainty the content of their speech, but *when* they need to know it. These cases underscore the idea that for every exercise of editorial discretion, there is a moment in which that discretion is actually exercised—a moment when an editor's judgment about what material to include or exclude in her compilation manifests as "speech" under the First Amendment. That moment is publication.

As Justice Powell observed in *Herbert v. Lando*, "whatever protection 'the exercise of editorial judgment' enjoys depends entirely on the protection the First Amendment accords the product of this judgment, namely, published speech."[141] The idea is both intuitive and a logical necessity. If editorial discretion requires that an editor know with certainty that her speech reflects her decisions about its contents, she must have a published version of that speech against which she can make the comparison—for instance, a printed book, a live broadcast or cable transmission, or a published website. Without it,

140. *Id.* This knowledge relates to the items included in the compilation—not knowledge as to the specific contents of each item within the compilation. *See Hurley*, 515 U.S. at 574 (finding general disapproval with an item's message sufficient to protect a speaker's decision to exclude it from a compilation). Separate from the speech inquiry (is something speech?) and protection inquiry (if so, is it *protected* speech?) discussed in Part IV, the knowledge as to the contents of each item within the compilation relates instead to the liability inquiry: Under what circumstances can one be held liable for *unprotected* speech? *See* Smith v. California, 361 U.S. 147, 154-55 (1959) (finding a bookseller cannot be liable for a book's illegal contents without some degree of knowledge as to its contents). *Smith*, decided in 1959, was decided before the editorial discretion doctrine was fully formed, but can be understood as an early editorial discretion case consistent with the principle of speech certainty. The bookseller's speech is the compilation of books. *Turner*, 512 U.S. at 675 (O'Connor, J., concurring in part and dissenting in part) ("Selecting which speech to retransmit is, as we know from the example of . . . bookstores, . . . no less communication than is creating the speech in the first place."). While the bookseller may not have known the contents of each book, he undoubtedly knew with certainty what books he did and did not procure and make available for sale in his shop. Thus, the bookseller's "speech" was characterized by speech certainty. And because he created the compilation of books, exercised judgment as to the contents of that compilation, and "published" the compilation by making the books available for sale, that speech is protected by the doctrine of editorial discretion. *See Moody*, 144 S. Ct. at 2402. The issue in *Smith* is not whether the bookseller's speech is protected; the Court is clear that it is. 361 U.S. at 150 ("[I]t . . . requires no elaboration that the free publication and dissemination of books . . . furnish very familiar applications of these constitutionally protected freedoms."). Instead, the issue is whether the bookseller can be held *liable* for unprotected third-party speech included in his protected compilation. *Id.* at 150-51.

141. 441 U.S. 153, 178 (1979) (Powell, J., concurring).

neither she nor the Courts could know whether the contents she "*decid[ed] . . .* [to] be included in or excluded from" her speech were in fact included in or excluded from it.[142]

Beyond intuitive and logical appeal, however, the publication requirement is reflected in the Court's analysis. As the *Herbert* Court recognized, the notion that the exercise of editorial discretion occurs at publication is deeply rooted in the editorial discretion cases. These cases uniformly concern efforts by the government "to control in advance the content of the *publication*" and "efforts to enjoin *publication* of specified materials."[143] It must have required (or sought to require) an editor to publish that which she meant to exclude, or to exclude that which she meant to publish. What the "publication" looks like varies for different types of compilations, such as newspapers (publication),[144] broadcasters (broadcast),[145] cable operators (transmission),[146] and parade organizers (the marching of the parade).[147] But the uniform feature of publication for all compilations is that the editorial decisions regarding their contents become final. It is for this reason that, in summarizing the editorial discretion cases, the *Moody* court expressly recognized that the "expressive

142. *Moody*, 144 S. Ct. at 2402 (emphasis added).

143. *Herbert*, 441 U.S. at 167-68 (emphasis added).

144. In the context of newspapers, the Court explicitly defined "the exercise of editorial control and judgment" as "[t]he choice of material to go into a newspaper." Miami Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 258 (1974). Justice White further emphasized that the doctrine's focus is on "decision[s] as to what copy will or will not be included in any given edition" of a newspaper. *Id.* at 261 (White, J., concurring). This definition makes it clear that for newspapers the proper unit of analysis is the contents contained in each new published edition of the newspaper.

145. For broadcasters, the focus is on content actually broadcast over the airwaves. *See, e.g.*, Columbia Broad. Sys., Inc. v. Democratic Nat'l Comm., 412 U.S. 94, 132 (1973) (discussing broadcasters' editorial discretion in terms of "air time").

146. For cable operators, the Court has zeroed in on "the total service offering to be extended to subscribers"—that is, what channels are included in its transmission. FCC v. Midwest Video Corp., 440 U.S. 689, 707-08 n.17 (1979). It concluded that First Amendment protection under the doctrine of editorial discretion attaches "[o]nce the cable operator has selected the programming sources [and] the cable system functions . . . as a conduit for the speech." Turner Broad. Sys., Inc. v. FCC, 512 U.S. 629 (1994).

147. For parades, the Court emphasized that "[t]he issue in this case is whether [the government] may require private citizens who organize a parade to include among the marchers a group imparting a message the organizers do not wish to convey." Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 559 (1995). Thus, the protected speech at issue in parades is the actual marching of the parade—its publication reflects the parade organizer's final decisions concerning which contingents to include and which to exclude.

activity" protected by the doctrine of editorial discretion is the "present[ation] of a curated compilation of speech."[148]

The Court's consistent and pervasive focus on publication throughout the editorial discretion cases reflects the First Amendment's speech certainty requirement: The editor must know with certainty the contents of her compilation at the moment she publishes it. Just as thoughts must manifest themselves as verbal, written, or printed speech to earn the protection of the First Amendment, so too must the editorial process result in the publication of a compilation of speech.[149] Only then does the editorial process result in any "speech" for the First Amendment to protect.

B. Expressive Conduct

Expressive conduct—another modern outgrowth of First Amendment doctrine—also relies on the speech certainty principle. Under this doctrine, wordless expression gains constitutional protection under the Speech Clause.[150] But claims that one's conduct is protected by the First Amendment have always been met with a healthy dose of suspicion. The Supreme Court, for example, has expressed that it "cannot accept the view that an apparently limitless variety of conduct can be labeled 'speech' whenever the person engaging in the conduct intends thereby to express an idea."[151] To that end, the Court developed what has become known as the *Spence* test—a two-prong analysis that polices the line between plain-old conduct and that which is expressive. For conduct to qualify as "speech" under the *Spence* test, (1) the speaker must have an "intent to convey a particularized message" and (2) there must be a great likelihood that the message will be understood by a reasonable observer.[152] For purposes of the speech certainty principle then, the conduct

148. Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2400 (2024); *see also id.* at 2402 ("[D]eciding on the . . . speech that will be included in or excluded from a compilation—*and then . . . presenting* the included items—is expressive activity of its own." (emphasis added)).

149. *See supra* Part I.A-.B; *see also Moody*, 144 S. Ct. at 2400 (finding the "expressive activity" protected by the doctrine of editorial discretion is the "present[ation] of a curated compilation of speech"); *Hurley*, 515 U.S. at 569-70 ("[A] private speaker does not forfeit constitutional protection simply by combining multifarious voices, or by failing to edit their themes to isolate an exact message as the exclusive subject matter *of the speech*. Nor, under our precedent, does First Amendment protection require a speaker to generate, as an original matter, each item *featured in the communication*." (emphasis added)). Interpreting the editorial discretion cases, Stuart Minor Benjamin reached a similar conclusion that editorial discretion contains a "communication requirement." Benjamin, *supra* note 112, at 1461 ("[I]n order to communicate, one must have a message that is sendable and receivable and that one actually chooses to send.").

150. Spence v. Washington, 418 U.S. 405, 409-10 (1974).

151. United States v. O'Brien, 391 U.S. 367, 376 (1968).

152. Texas v. Johnson, 491 U.S. 397, 404 (1989) (quoting *Spence*, 418 U.S. at 410-11).

that communicates the message is the purported "speech." Therefore, we analyze the speech certainty principle in these cases not with respect to whether the speaker is certain of the words that she speaks, but rather of the acts she carries out.

This Subpart will show that both prongs of the *Spence* test assume the speech certainty principle—that the speaker knew what her conduct was at the moment she performed it. As for the first prong, a speaker only had the requisite intent to convey a message through some conduct if she knew what that conduct was. And for the second prong, an observer's understanding of that message requires that the speaker did, in fact, perform the intended conduct. Once the conduct is performed, it is undoubtedly characterized by speech certainty. It is impossible, then, for conduct to satisfy the *Spence* test without also satisfying the speech certainty principle.

1. Intent to convey a particularized message through the conduct[153]

Starting with perhaps the most obvious point, the expressive conduct line of cases has always assumed that the speaker knows what she is doing and saying through her conduct; that is, she is certain of the contents of her speech. That simple fact is evident from the first prong of the expressive conduct test: that the speaker intended to communicate a message with her conduct.[154] For an action to qualify as expressive conduct, the speaker must be "intimately connected with the communication advanced."[155] To that end, the message must be "intentional"[156] and the conduct must be "sufficiently imbued with the

153. Some scholars have argued that the first prong of the *Spence* analysis was weakened in *Hurley*. *See, e.g.*, Sandy Tomasik, *Can You Understand This Message? An Examination of* Hurley v. Irish-American Gay, Lesbian & Bisexual Group of Boston*'s Impact on* Spence v. Washington, 89 ST. JOHN'S L. REV. 265, 267 (2015) ("[*Hurley*] potentially altered [the *Spence*] test."). Indeed, some Circuits have done away with the first prong entirely. *See, e.g.*, Holloman *ex rel.* Holloman v. Harland, 370 F.3d 1252, 1270 (11th Cir. 2004) ("Thus, in determining whether conduct is expressive, we ask whether the reasonable person would interpret it as *some* sort of message, not whether an observer would necessarily infer a *specific* message."). But because some Circuits still apply the first prong, and the Supreme Court's follow-on case in seems to have revived it, we address its application in full. *See, e.g.*, Blau v. Fort Thomas Pub. Sch. Dist., 401 F.3d 381, 388 (6th Cir. 2005) ("Claimants must show that their conduct 'conveys a particularized message.'" (quoting *Spence*, 418 U.S. at 411)); Rumsfeld v. F. for Acad. & Institutional Rts., Inc., 547 U.S. 47, 65-66 (2006).

154. *Johnson*, 491 U.S. at 404.

155. Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 576 (1995) ("[W]hen dissemination of a view contrary to one's own is forced upon a speaker intimately connected with the communication advanced, the speaker's right to autonomy over the message is compromised.").

156. *Johnson*, 491 U.S. at 406; *Spence*, 418 U.S. at 410-11.

elements of communication" to gain the protection of the First Amendment.[157] Therefore, the intended message and the conduct used to communicate it are inextricably interconnected. If a speaker has an intent to convey a message through some conduct, that intent is only evinced when she actually and knowingly performs that conduct. The performance of the conduct—the "speech"—thus reveals the *Spence* test's commitment to the speech certainty principle.

While the speaker's action might not communicate a *particularized* message, the Supreme Court has required particularized *conduct* before expressive conduct protection will attach.[158] And that conduct must transmit whatever message the speaker intends to express. In *Texas v. Johnson* for example, the Supreme Court found it critical and self-evident that the "expressive, overtly political nature of [the speaker's] conduct was . . . intentional."[159] Indeed, Johnson intended to communicate an anti-war message by burning the flag, and his message was only transmitted when he actually set fire to the Stars and Stripes.[160] The intentional message is thus only communicated through the intentional performance of the conduct. And, at risk of stating the obvious, an intentional performance requires that the performer know what she performed when she did.

The same is true for the Tinkers in their suit against the Des Moines Independent Community School District.[161] There, several young students in a public school were punished for wearing black armbands in the classroom to express their disapproval of the Vietnam War. But importantly, the children's anti-war message was only communicated the moment they donned a black armband.[162] Without either the intended anti-war message or the intentional decision to wear the armband symbolizing that expression, there wouldn't have been any "speech" to protect. And without that certain conduct, the intent isn't—indeed, can't be—communicated. Thus, the speech certainty principle isn't a departure from pre-existing doctrine; it's baked into its very foundation.

2. Great likelihood that the message will be understood by a reasonable observer

The speech certainty principle's application to the second prong—the great likelihood that a reasonable observer will understand the message—follows

157. *Spence*, 418 U.S. at 409.

158. *See Johnson*, 491 U.S. at 406.

159. *Id.*

160. *Id.*

161. Tinker v. Des Moines Indep. Cmty. Sch. Dist., 393 U.S. 503 (1969).

162. *Id.* at 504.

directly from the first.[163] Time and time again, the Court has emphasized that the speaker's message must "be understood by those who view[] it."[164] But the Supreme Court has imposed a high bar on this second prong: the message that is expressed through the conduct must be "overwhelmingly apparent."[165] If an observer could reasonably interpret the conduct as expressing some other kind of message, then the conduct isn't entitled to protection under the Speech Clause.

In effect, this means that the speech certainty principle is a prerequisite to this more demanding second prong of the *Spence* test. While a speaker may intend to communicate a particularized message via some conduct, and in fact execute that conduct, it may nonetheless fall short of qualifying as *expressive* conduct for reasons unrelated to the speaker's certainty in their purported "speech."

In *Rumsfeld v. Forum for Academic & Institutional Rights, Inc.*, for example, the Court considered whether a federal requirement that law schools allow military recruiters onto their campuses constituted compelled speech.[166] The Circuit Court had held that the law schools' "speech" in this case was the expressive act of rejecting military recruitment.[167] To force the schools to welcome military recruiters on campus, the schools contended, would amount to compelling an implicit endorsement of the military and its actions.[168] But the Supreme Court rejected their theory on the ground that a reasonable observer could take away any number of messages from the fact that military recruiters didn't have a presence on law school campuses:[169]

> An observer who sees military recruiters interviewing away from the law school has no way of knowing whether the law school is expressing its disapproval of the military, all the law school's interview rooms are full, or the military recruiters decided for reasons of their own that they would rather interview someplace else.[170]

In other words, the law school did in fact bar military recruiters from its campus, and knew it did so at the moment it made the decision, thus satisfying the speech certainty principle. But because a reasonable observer couldn't have

163. *See Johnson*, 491 U.S. at 404.

164. *See, e.g.*, Spence v. Washington, 418 U.S. 405, 411 (1974).

165. *Johnson*, 491 U.S. at 406; *see also* Rumsfeld v. F. for Acad. & Institutional Rts., Inc., 547 U.S. 47, 66 (2006) ("For example, the point of requiring military interviews to be conducted on the undergraduate campus is not 'overwhelmingly apparent.'" (quoting *Johnson*, 491 U.S. at 406)).

166. *Rumsfeld*, 547 U.S. at 60-61.

167. *Id.*

168. *Id.* at 64-65.

169. *Id.*

170. *Id.* at 66.

understood the law school's message from its conduct alone, the Court held that the schools weren't engaged in speech-protected activity.

Underlying this interconnected nature of the message and the conduct is the obvious necessity that the speaker knows what she's saying when she says it—or here, that she knows what she's doing when he engages in the conduct. Unless the speaker executes her conduct with the requisite certainty in her actions, there's no chance that the reasonable observer will understand the intended message. Indeed, it would be absurd to assume otherwise. The speaker, at a minimum, must know the conduct she engaged in for an observer to understand the communicative content underlying that conduct. Otherwise, the actor doesn't "speak" at all, and her conduct is relegated to non-expressive acts unworthy of First Amendment protection.[171]

The *Spence* test requires that a speaker's message and her expressive conduct are fully intertwined in order to gain the First Amendment's protection. This makes eminent sense given the Supreme Court's general skepticism towards those that may try to shield otherwise routine impermissible conduct from criminal sanction by claiming that it's their speech. But part and parcel with those requirements is also the requirement that the speaker had the requisite intent to communicate her message through conduct that she is certain she carried out, making the speech certainty principle a necessary element underlying the expressive conduct analysis.

III. Understanding Algorithmic Output

Until roughly the last decade, the principle of speech certainty was so inherent to speech that articulating its existence was never necessary. It underpinned the spoken, written, and printed word, broadcast and cable transmissions, and, as we will explain in this part and the next, the delivery of

171. When a message isn't "overwhelmingly apparent" from the conduct and context alone, some litigants have nonetheless tried to shoehorn the conduct into the ambit of the First Amendment with explanatory speech. Such speech, they say, shores up the connection between the message and the conduct. The Supreme Court has rejected such attempts, instead holding that the conduct must "speak" for itself. If the "expressive component" of the actions "is not created by the conduct itself, but by the speech that accompanies it," then it doesn't qualify as expressive conduct. *Rumsfeld*, 547 U.S. at 66. In fact, "[t]he fact that such explanatory speech is necessary is strong evidence that the conduct at issue here is not so inherently expressive that it warrants protection." *Id.* Indeed, "[i]f combining speech and conduct were enough to create expressive conduct, a regulated party could always transform conduct into 'speech' simply by talking about it" thereby eviscerating the line between speech and conduct. *Id.* Post-hoc or even concurrent explanations of what one means to say aren't enough if the conduct doesn't get the job done on its own. In other words, it's the conduct that must do the talking, and the only way that the conduct can do that is if the speaker is certain of what that conduct was.

content via code on the early internet. All communication was characterized by speech certainty; it was simply impossible to speak without knowing what you said when you said it.

But that has now changed with programmers, platforms, and artificial intelligence companies claiming the output of machine learning algorithms as their speech.[172] While traditional algorithms faithfully follow the rules written by a programmer to determine their output, machine learning algorithms write their own rules. And these rules calculate probabilities to make predictions.[173] For example, based on the combination of words in a post published by a particular user, what is the likelihood that the post includes a claim that a violent tragedy did not occur? Or based on the way the pixels in an image are arranged, that the image includes a derogatory sexualized photoshop?[174] But the nature of these probabilities means that they can never be 100% certain in the accuracy of their output.[175] Neither, then, can their programmers be certain of the contents of that output.

In this Part, we put the speech certainty analysis on hold as we look under the hood to explain how machine learning algorithms work. Our goal is to provide lawyers, policymakers, and advocates with a baseline understanding of the technology underpinning machine learning and artificial intelligence—and to illustrate how fundamentally it differs from what we typically understand computer programming to be. For our purposes, two fundamental differences emerge: (1) the machine learning programmer does not write the rules that govern his algorithm and (2) the machine learning programmer cannot explain how his algorithms work. These two facts prove decisive in Part IV when we analyze the speech certainty of machine learning models.

A. What We Mean By "Machine Learning"

In writing this section, we have relied heavily on the introductory Machine Learning Crash Course developed by Google.[176] We do so in the

172. *See supra* note 1 (citing case briefs in which platforms have claimed First Amendment protection for their algorithmic output).

173. *See generally* KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE, at xxvii (2012) (explaining the relationship between machine learning and probability theory).

174. *See, e.g.*, *Adult Sexual Exploitation*, META, https://perma.cc/977X-C2SY (archived Oct. 20, 2024).

175. *See* Loukides, *supra* note 4.

176. *Machine Learning Crash Course with TensorFlow APIs*, GOOGLE FOR DEVELOPERS https://perma.cc/Y66Q-CTVE (archived Oct. 20, 2024). Much to the authors' chagrin, Google revised its Machine Learning Crash Course between the writing of this Article and its publication. *See* Sanders Kleinfeld, *Our Machine Learning Crash Course Goes in Depth on Generative AI*, GOOGLE KEYWORD (Nov. 12, 2024), https://perma.cc/DMB6-

footnote continued on next page

belief that the baseline understanding provided to programmers by Google—a pioneering and leading practitioner of machine learning[177]—ought to suffice as a baseline understanding for non-technical lawyers, policymakers, and advocates as well.[178]

In relying on Google's course, this Part focuses on one of the most popular approaches to machine learning: supervised machine learning algorithms that rely on a mathematical process called gradient descent.[179] (Admittedly, that's a mouthful, so just as Google uses the shorthand "machine learning" throughout its course to refer to this method, so do we throughout this section.) Although there are several schools of machine learning, and many different methods within those schools,[180] we limit this explainer to supervised machine learning models that rely on gradient descent. We do so for two reasons.

First, gradient descent is a primary method by which machine learning models write their own rules. It powers many of the most common approaches of machine learning.[181] These common approaches underpin the algorithms behind content moderation on social media platforms, which are the focus of this Article.[182] Understanding gradient descent is therefore critical to understanding why the output of these content moderation algorithms is not the platforms' speech.

Second, the narrow conclusion we draw from *supervised* models with gradient descent can be generalized across other forms of machine learning

---

YGRT. The Google update did not substantively alter the content of the courses such that it affects how the ideas in the Article are discussed, but significantly reorganized the curriculum. In this Article, the authors and editors have cited to perma.cc links that contain the archived webpages as they were used in the drafting of this Article.

177. *Machine Intelligence*, GOOGLE RSCH., https://perma.cc/P465-JS96 (archived Oct. 20, 2024) ("Google is at the forefront of innovation in Machine Intelligence, with active research exploring virtually all aspects of machine learning, including deep learning and more classical algorithms.").

178. Just as programmers can move on to advanced courses to improve their understanding, so too should readers of this piece who seek to deepen their knowledge. For the purposes of this Article, however, the Machine Learning Crash Course is sufficient.

179. *See infra* Part III.B.2.c.

180. *See generally* DOMINGOS, *supra* note 2, at xvii (explaining the differences between the various schools of machine learning).

181. IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 96-97 (2016) ("Most deep learning algorithms are based on an optimization algorithm called stochastic gradient descent.").

182. Singh, *supra* note 6 ("In response to growing global pressure from governments and the public to take down violating content quickly, Facebook has invested heavily in automated tools for content moderation."); Charlotte Jee, *This is How Facebook's AI Looks for Bad Stuff*, MIT TECH. REV. (Nov. 29, 2019), https://perma.cc/CZF9-4W9A ("The vast majority of Facebook's moderation is now done automatically by the company's machine-learning systems.").

that also rely on gradient descent. That is, if the output of supervised machine learning with gradient descent lacks speech certainty—and therefore falls outside the First Amendment's protection—the output of all machine learning relying on gradient descent does too. This is because supervised machine learning tends to be more directed by humans than the other methods of machine learning.[183] And as the programmer's involvement in a machine learning model diminishes, the argument that it lacks speech certainty only gets stronger. If supervised machine learning with gradient descent isn't speech, then no machine learning with gradient descent can be either.

With that prelude, let's explain how these machine learning models work.

An Opportunity for Technophobic Readers

A modest word of caution to readers averse to getting in the technical weeds: Machine learning is technically complicated. And to explain how it works, we cannot and do not shy away from discussion of technical concepts. We want readers to see what machine learning programmers are doing—and what they are not doing—to make it clear that they are not simply writing code. For readers who are more interested in the legal implications of that conclusion rather than understanding the facts that compel it, you are welcome to skip the rest of this section so long as you are willing to accept the premises listed below as true. With these key takeaways, the rest of our argument should flow quite naturally from Part IV onward.

1. Traditional code executes only the rules the programmer has written. With traditional programming, the programmer determines the rules the algorithm should follow and translates those rules into code. When the code runs, the programmer can know with complete certainty that it will execute the rules as written.[184]

2. Machine learning algorithms write their own rules to make predictions. The only function of these algorithms is to make predictions. They do so by writing their own rules using complex math that gets executed automatically. Programmers' primary role in the process is to tell the algorithm what it should be predicting (e.g., the likelihood that a given email is spam), to determine what variables to consider (e.g., who sent the email? At what time of day?), to make sure that the data fed into the model is reliable, and to adjust certain "knobs" that facilitate the model's automatic

183. Rachel Wilka, Rachel Landy & Scott A. McKinney, *How Machines Learn: Where Do Companies Get Data for Machine Learning and What Licenses Do They Need?*, 13 WASH. J.L. TECH. & ARTS 217, 222-25 (2018) (explaining the three major categories of machine learning: supervised, unsupervised, and reinforcement).

184. *See infra* Parts III.B, III.B.1.

math. This work is harder and more complicated than it sounds. But however challenging the work, programmers don't participate in the algorithm's fundamental task: determining the logic that shapes the algorithm's predictions. That is, the machine independently decides how much each variable should matter in its predictions and how each variable influences the prediction.[185]

3. Programmers cannot explain why machine learning algorithms make the predictions they do. In most cases, the algorithms are "black boxes"—rules so complex that they are incapable of human understanding. In others, they are the product of too large a set of rules for any human to practically process. This inherent opacity, coupled with the fact that the programmers did not write the rules, means that programmers generally do not understand how they work.[186]

4. The output of machine learning algorithms will always contain errors that cannot be attributed to the programmer. Machine learning algorithms calculate probabilities to make predictions. The nature of probability is that the algorithms can never be 100% certain in any outcome, and as a result, they will inevitably be wrong at least some of the time. To be sure, programmers can use traditional code to make predictions too—predictions that also inevitably get things wrong. But with traditional code, the programmer writes the rules, so even when the predictions are wrong, the algorithm has faithfully executed the rules written by the programmer; the errors can be directly attributed to him. With machine learning, the programmer neither writes the algorithm's rules, nor can he fully explain them. Thus, when the algorithm makes a mistake, it cannot be attributed to the programmer.[187]

B. Code: The Shift from Traditional Programming to Machine Learning

"Code is not constant," Lawrence Lessig wrote in 1999.[188] "It changes. . . . How it changes depends on the code writers."[189] At the turn of the twenty-first century, this was a truism. Code could not change without the intervention of a code writer any more than the words on a page could change without the intervention of an editor. In this traditional mode of programming, a code

185. *See infra* Part III.B.2.c.

186. *See infra* Part III.B.2.e.

187. *See infra* Part III.B.2.f.

188. LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 109 (1999).

189. *Id.*

writer—we refer to them as programmers—writes a set of rules and then, when the program runs, it executes those rules according to the will of the programmer.[190] These rules are expressed in code,[191] which is "intended to express each idea completely unambiguously, so that each program does exactly one, completely predictable thing."[192] Code is what makes the software inside everything from iPhones to dishwashers work. And as venture capitalist Marc Andreesen famously wrote in 2011—an era we can now see as the twilight of the traditional programming paradigm—"[s]oftware is eating the world."[193]

This Article's narrow focus is on the use of code to make predictions, which have become a surprisingly central feature of the software we use every day. Just as software ate the world a decade ago, predictions are now eating

190. "The 'classical stack' of Software 1.0 is what we're all familiar with—it is written in languages such as Python, C++, etc. It consists of explicit instructions to the computer written by a programmer. By writing each line of code, the programmer identifies a specific point in program space with some desirable behavior." *See* Andrej Karpathy, *Software 2.0*, MEDIUM (Nov. 11, 2017), https://perma.cc/A6MX-VWDH. Andrej Karpathy is a computer scientist and educator who served as the Senior Director of AI at Tesla. ANDREJ KARPATHY, https://perma.cc/XJW5-Q2RV (archived Nov. 18, 2024).

191. Code can be written in a variety of programming languages, but they all fundamentally work the same way. *See* JAMES GRIMMELMANN, INTERNET LAW: CASES AND PROBLEMS 24 (2014). "[C]onsider the process of averaging two numbers. If asked to describe averaging, you might say '[a]dd the numbers together, and then take half of the result.' This is an *algorithm*, a step-by-step process for carrying out a calculation." *Id.* at 25. In code, the algorithm looks like this:

Python:	C:	Scheme:

def average (x y):
  sum = x + y;
  return sum / 2;
	int average (int x, int y)
	  {
	  int sum;
	  sum = x + y;
	  return sum / 2;
	  }
		(define average
		  (lambda (x y)
		  (/ (+ x y) 2)))

Id. at 25-26.

192. *Id.* at 25 ("A computer programmer's job consists of translating informal descriptions . . . into a sufficiently precise series of statements that a computer could execute them.").

193. *See also* Alicia Solow-Niederman, *Emerging Digital Technology and the "Law of the Horse"*, UCLA L. REV. (Feb. 19, 2019), https://perma.cc/UEH9-6BQ2 ("[I]t is increasingly difficult to think of a sector or domain that is *not* affected by code."); James Somers, *A Coder Considers the Waning Days of the Craft*, NEW YORKER (Nov. 13, 2023), https://perma.cc/YXS8-47CC ("Bodies of knowledge and skills that have traditionally taken lifetimes to master are being swallowed at a gulp. Coding has always felt to me like an endlessly deep and rich domain. Now I find myself wanting to write a eulogy for it."). *See generally* Marc Andreesen, *Why Software Is Eating the World*, ANDREESEN HOROWITZ (Aug. 20, 2011), https://perma.cc/7JA7-R7YY (describing software's impact across industries).

software.[194] When we search for something, Google and Bing predict the websites most likely to give us the information we're looking for.[195] When we open TikTok, YouTube, Instagram, Facebook, or X/Twitter, they predict the posts and videos that we're most likely to enjoy.[196] And when we ask a chatbot a question, it predicts the response most likely to give us the answer we're looking for.[197] The reality is that when we talk about algorithms—be it in terms of "surveillance capitalism" or the "tyranny of Big Tech"—what we're talking about is predictions.[198]

How predictions are made using code, however, has undergone a massive transformation in the last ten to fifteen years, the details of which have not yet been widely recognized.[199] The fundamental feature of code—that it executes specific instructions written by a programmer—used to apply to predictions.[200] But as predictions have increasingly become the province of machine learning in recent years, this truism no longer applies.[201] Pedro Domingos, a machine

194. *See* Karpathy, *supra* note 190.

195. *See* Pandu Nayak, *Understanding Searches Better Than Ever Before*, GOOGLE: KEYWORD (Oct. 25, 2019), https://perma.cc/RK6S-Y2G8 (Google); *Introducing the Next Wave of AI at Scale Innovations in Bing*, MICROSOFT BING BLOGS (Sept. 23, 2020), https://perma.cc/MW3C-RD3B (Bing).

196. Arvind Narayanan, *TikTok's Secret Sauce*, KNIGHT FIRST AMEND. INST. (Dec. 15, 2022), https://perma.cc/U8NR-25XR (TikTok); Goodrow, *supra* note 10 (YouTube); Twitter, *Twitter's Recommendation Algorithm*, X ENGINEERING (Mar. 31, 2023), https://perma.cc/RXV7-KUDJ (X/Twitter).

197. *See, e.g.*, *OpenAI's Technology Explained*, OPENAI (Oct. 11, 2023), https://perma.cc/CHF7-X2KL ("We teach the model to respond in ways that people find more useful, and to decline to respond in ways that we believe would be harmful.").

198. John Laidler, *High Tech Is Watching You,* HARVARD GAZETTE (Mar. 4, 2019), https://perma.cc/P7Y2-82CB (defining "surveillance capitalism as the unilateral claiming of private human experience as free raw material for translation into behavioral data . . . packaged as prediction products"); *cf.* HAWLEY, *supra* note 13, at 4-5 (2021) (discussing how large technology corporations use data and predictive algorithms to "manipulate individuals to change their behavior").

199. ARVIND NARAYANAN, UNDERSTANDING SOCIAL MEDIA RECOMMENDATION ALGORITHMS 24-25 (2023), https://perma.cc/QT8Z-YUUC (noting for social media platforms, recommendation algorithms "only started happening in the 2010s"); ANDREW NG, MACHINE LEARNING YEARNING 10 (2018) (noting that the two biggest drivers of recent progress in deep learning are data availability and computational scale); Nicolas Koumchatzky & Anton Andryeyev, *Using Deep Learning at Scale in Twitter's Timelines*, X ENGINEERING (May 9, 2017), https://perma.cc/V9S8-CVY6 ("In the field of machine learning, deep learning and the development of AI-related work these last few years has led to an unprecedented (and ongoing) burgeoning of new ideas and algorithms."); Goodrow, *supra* note 10.

200. *See, e.g.*, NARAYANAN, *supra* note 199, at 29-30 (describing how some predictions were "manually programmed" before being "replaced by machine learning").

201. *See* David Auerbach, *The Programs That Become the Programmers*, SLATE (Sept. 25, 2015), https://perma.cc/VD6E-WKXE (discussing the "significant change from the traditional programming paradigm").

learning pioneer, summarizes it succinctly in his book *The Master Algorithm*: "With machine learning, computers write their own programs, so we don't have to."[202]

1. Predictions with traditional programming

The development of computer programming in the nineteenth and twentieth centuries allowed people to use code to automate chains of logical reasoning.[203] "Believe it or not," Domingos writes, "every algorithm, no matter how complex, can be reduced to just these three operations: AND, OR, and NOT. . . . By combining many such operations, we can carry out very elaborate chains of logical reasoning."[204] Because predictions are simply a form of logical reasoning, they can be expressed in code. "Without machine learning," Google explains, "programmers must manually write instructions to make useful predictions."[205] And until the advent of machine learning, that's precisely what programmers did.[206]

Consider Facebook's EdgeRank algorithm circa 2010, prior to the company's widespread adoption of machine learning.[207] EdgeRank determined how posts in a user's Newsfeed would be ranked. To do that, Facebook's engineers wrote a formula that would analyze "every item that could potentially be shown to the user" to predict the likelihood that a user would engage with it.[208] The formula then ranked these posts from those the user was deemed most likely to engage with to those least likely.[209] As Arvind

202. DOMINGOS, *supra* note 2, at 6; *see also* Karpathy, *Software 2.0*, *supra* note 190 ("Software (1.0) is eating the world, and now AI (Software 2.0) is eating software.").

203. *See generally* David Hemmendinger, *Computer Programming Language*, BRITANNICA, https://perma.cc/4W8Y-CE8J (archived Oct. 20, 2024) (identifying various programming languages).

204. Domingos, *supra* note 2, at 2.

205. Google Open Online Education, *What Is ML?*, YOUTUBE, at 0:11 (Feb. 24, 2022), https://perma.cc/YMQ7-3UVW.

206. NARAYANAN, *supra* note 199, at 29 ("Although the approach of optimization based on machine learning is ubiquitous today, it wasn't always the case."); *Why Google Went from a Rules-Based to a ML-Based Search Engine*, ASK THE SEARCH ENGINEER, https://perma.cc/RRW5-4DVB (archived Oct. 20, 2024) (discussing the history of Google's shift from a rules-based search engines to a machine learning-based search engine).

207. *See* Jeff Widman, *EDGERANK*, https://perma.cc/3MVE-RUH5 (archived Oct. 20, 2024); Jeffrey Dunn, *Introducing FBLearner Flow: Facebook's AI Backbone*, ENGINEERING AT META (May 9, 2016), https://perma.cc/4LU4-J938 ("In late 2014, we set out to redefine machine learning platforms at Facebook from the ground up, and to put state-of-the-art algorithms in AI and ML at the fingertips of every Facebook engineer.").

208. NARAYANAN, *supra* note 199, at 29.

209. *See* Widman, *supra* note 207.

Narayanan, a professor of computer science at Princeton, explains, EdgeRank's two most important variables were:

- An "affinity score" representing "how much the user in question wants to see updates from the poster. This was . . . a manually programmed formula . . . ; no machine learning was involved."[210]
- An "item type weight," that identified whether the post included text, photo, or video and "reflected Facebook engineers' predictions regarding the type of content that was more engaging. . . . These were also manually set . . . ."[211]

Naranayan's key point about EdgeRank is that its "two key ingredients" were manually programmed to reflect the judgments of Facebook's engineers. That is, the programmers wrote the algorithm in code. In the traditional programming paradigm, they had no other choice; manual programming was simply how algorithms were created.

Domingos explains this traditional programming paradigm clearly: "Every algorithm has an input and an output: the data goes into the computer, the algorithm does what it will with it, and out comes the result."[212] EdgeRank is a quintessential example: Facebook's vast collection of user posts provides the data; the algorithm written in code by Facebook's programmers determines each post's desired rank in a given user's Newsfeed; and the output is a user's Newsfeed ranking those posts accordingly. Should the programmers change their minds about how the algorithm worked—what variables it took into account and how much they should matter—they would have to update it the old-fashioned way: by manually editing the code.[213] For programmers of the early web in the 1990s and 2000s, this was the only way.[214] To make predictions with traditional programming, the programmer must write the rules that underpin the predictions.

---

210. NARAYANAN, *supra* note 199, at 29 ("The two key ingredients in the formula are the affinity score and the item type weights.").

211. *Id.*

212. DOMINGOS, *supra* note 2, at 6.

213. NARAYANAN, *supra* note 199, at 29 (explaining that EdgeRank's variables were "manually set").

214. Google Open Online Education, *supra* note 205, at 0:11 ("Without machine learning, programmers must manually write instructions to make useful predictions."). *Compare, e.g.*, Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 30 COMPUT. NETWORKS & ISDN NETWORKS 107, 111-15 (1998) (describing Google's early algorithm, PageRank), *with* Matthew Richardson, Amit Prakash & Eric Brill, *Beyond PageRank: Machine Learning for Static Ranking*, 2006 PROC. 15TH INT'L CONF. WORLD WIDE WEB 707, 707-08, https://perma.cc/67TQ-AK2J (describing PageRank as a "static . . . ordering of web pages" and differentiating it from a proposed method based on machine learning).

2. Predictions with machine learning

Machine learning has ushered in a paradigm shift for how programmers make predictions. Whereas a traditional programmer writes rules that tell an algorithm how to analyze data to generate a prediction, "[m]achine learning turns this around."[215] The machine learning programmer tells the algorithm what he wants it wants to predict, provides the algorithm with data, and the algorithm determines for itself the rules that generate the predictions.[216]

"By building [machine] learning systems, we don't have to write these rules anymore," explained John Giannandrea, Apple's senior vice president of Machine Learning and AI Strategy and former head of search and artificial intelligence at Google.[217] One might be tempted to write off such statements as the overpromising bluster of the Silicon Valley hype cycle.[218] But fanciful as it may sound, this Subpart explains that it's true.[219]

The remainder of this Subpart explains three fundamental characteristics of machine learning: (1) machine learning algorithms write their own rules; (2) machine learning programmers cannot explain those rules; and (3) the algorithms' predictions are guaranteed to be wrong at least some of the time. Together, as we explore in Part IV, these characteristics prove decisive in the First Amendment analysis.[220]

---

215. DOMINGOS, *supra* note 2, at 6.

216. *See Model*, GOOGLE FOR DEVELOPERS: MACHINE LEARNING GLOSSARY, https://perma.cc/4LLP-UTDE (archived Oct. 20, 2024) ("A human programmer codes a programming function manually. In contrast, a machine learning model gradually learns the optimal parameters during automated training.").

217. Craig G. Karl, *AI Is Transforming Google Search. The Rest of the Web Is Next*, WIRED (Feb. 4, 2016), https://perma.cc/SDT4-66GW ("Increasingly, we're discovering that if we can learn things rather than writing code, we can scale these things much better."); *Apple Leadership: John Giannandrea*, APPLE, https://perma.cc/U6F2-979Z (archived Oct. 20, 2024).

218. *Gartner Hype Cycle*, GARTNER, https://perma.cc/9P3F-JAVW (archived Dec. 12, 2024).

219. But do not mistake this to mean that the programmer is irrelevant to the overall process, or that his work is easy. Machine learning programmers have challenging jobs that demand expertise and painstaking labor. It's just that the work entailed in that labor looks radically different from what we have come to expect from traditional programming. The difference is the degree of control. Rather than deciding on the precise logic for the algorithm, as a traditional programmer would do in code, the machine learning programmer facilitates the model's ability to determine the precise logic via gradient descent. This facilitating role, however, leaves the critical act—determining the logic—to the model. For more on the programmer's role in machine learning, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669-701 (2017) (explaining the stages of machine learning).

220. *See infra* Part IV.B.3.c.

a. Machine learning: Key terms

Google defines machine learning as "systems [that] learn how to combine input to produce useful predictions on never-before-seen data."[221] Within the field of machine learning, the most common method is known as supervised learning.[222] Supervised machine learning works by "[t]raining a *model* from *features* and their corresponding *labels*."[223] We define each of these three foundational terms in more detail below.

- *Label*. "A label is the thing we're predicting."[224] And in supervised machine learning, the programmer creating the model determines what that is. To use an example, "in a spam detection dataset, the label would probably be either 'spam' or 'not spam.'"[225]
- *Feature*. A feature is "an input variable" associated with the prediction we're trying to make.[226] Again, in supervised machine learning, the programmer determines the variables that the model should consider. "In the spam detector example, the features could include the following: words in the email text, sender's address, time of day the email was sent, email contains the phrase 'one weird trick.'"[227] Machine learning projects can include anywhere from a single feature to millions of features.[228]
- *Model*. The "model defines the relationship between features and label."[229] That is, the model calculates the influence that a given feature plays in determining the proper label. In the spam detector example, "a spam detection model might associate certain features strongly with 'spam.'"[230]

The life of a machine learning model includes two phases: (1) training, in which the model is created based on data provided to it by the programmer. It's in the training phase that the model writes its own rules. And (2) inference, in

221. *Framing: Key ML Terminology*, GOOGLE FOR DEVELOPERS, https://perma.cc/9TJ6-PTA5 (archived July 31, 2024).

222. NG, *supra* note 199, at 9 ("There are many forms of machine learning, but the majority of Machine Learning's practical value today comes from supervised learning."); *Supervised Learning*, GOOGLE FOR DEVELOPERS, https://perma.cc/Z5M3-HEU7 (archived May 28, 2023) ("Supervised learning is the dominant ML system at Google.").

223. *Supervised Machine Learning*, *supra* note 216 (emphasis added).

224. *Framing: Key ML Terminology*, *supra* note 221.

225. *Label*, *supra* note 216.

226. *Framing: Key ML Terminology*, *supra* note 221.

227. *Id.*

228. *See id.*

229. *Id.*

230. *Id.*

which the model is applied to never-before-seen data. During the inference phase, the model applies its rules to generate output.[231]

b. Training the model: Labeled examples & parameters

The goal of training the model is "to work out the best solution for predicting the labels from the features."[232] In other words, the goal is to figure out the rules that make the predictions. Programmers begin this process by "show[ing] the model *labeled examples* and enabl[ing] the model to gradually learn the relationships between features and label," which are known as *parameters*.[233]

A labeled example is an exemplar of the thing a programmer wants the model to predict, containing one or more features and manually tagged with the accurate label.[234] In the spam detector example, a labeled example might be an actual spam email that the programmer has identified and labeled as spam, and includes the following features: "sent from an email address not in your contacts," "sent at 2:43 A.M.," and "includes the words 'one weird trick.'" Other labeled examples may be emails including those same features, but identified as "not spam," or emails marked "not spam" that include some, but not all of those features.

The process of creating labeled examples is where programmers spend most of their time in the creation of a machine learning model[235]—and this marks the first departure from traditional programming. In traditional programming, a programmer might estimate that an email coming from an unknown sender is, say, 6% likely to be spam. He would come up with similar calculations for all sorts of other features, develop a mathematical formula that generates a sufficiently satisfying overall estimate of whether a given email is spam, translate that formula into code, and tell the computer to run it on every incoming email to determine whether it should make it into the user's inbox or be placed in the spam folder.

That traditional process is definitively not what programmers are doing when they train a machine learning model to predict whether an email is likely to be spam. Instead, machine learning programmers painstakingly label data and feed those labeled examples into the model.[236] And because the model

231. *Weight*, *supra* note 216 ("Training is the process of determining a model's ideal weights; inference is the process of using those learned weights to make predictions.").

232. *Supervised Learning*, GOOGLE FOR DEVELOPERS, https://perma.cc/S4GM-NAWL (archived Oct. 20, 2024) (emphasis added).

233. *Framing: Key ML Terminology*, *supra* note 221 (emphasis added).

234. Lehr & Ohm, *supra* note 219, at 673-76.

235. Karpathy, *supra* note 190.

236. *Id.*

is only as good as the data it's trained on, machine learning programmers spend most of their time "curating, growing, massaging[,] and cleaning labeled datasets."[237]

From there, the model determines the relationships between the features and the label on its own. These relationships are known as *parameters*—the elements of the mathematical formula that describe how much influence each variable has on the prediction.[238] As a basic example, if there's a 15% probability that an email coming from an unknown sender is spam, the 15% figure would be the parameter that describes the relationship between the feature (does the email come from an unknown sender?) and its role in predicting the correct label (is the email spam or not spam?).

At its core, training a model is the process of determining the ideal parameters comprising a model based on an analysis of labeled examples.[239] The next Subpart explains how it does so through a process called gradient descent—the critical step in the process by which the model writes its own rules.

c. Writing the rules: Gradient descent

Gradient descent is among the central processes underpinning today's machine learning revolution.[240] And because it is the process by which machine learning models write their own rules—and how they do so in a way the programmer cannot explain—gradient descent is the key to understanding why such models merit distinct treatment under the First Amendment. Thus, in this Subpart, we explain how gradient descent works to show that it, independently of the programmer, writes the rules underpinning machine learning models' predictions.

Whether a prediction is made by gradient descent, a programmer, or a shaman, it is ultimately judged by its accuracy. In machine learning parlance, a model's accuracy is measured by its *loss*—the gap between its prediction and the right answer.[241] A model's goal, then, is to reduce loss in order to improve the accuracy of its predictions. And because, as explained above, the accuracy of a model's predictions is determined by the values of the parameters it has selected, a machine learning model reduces loss by refining the values of those

237. *Id.*

238. *See Parameter*, *supra* note 216; *Weight*, *supra* note 216.

239. *Training*, *supra* note 216.

240. *See* Daniel Godoy, *Gradient Descent, the Learning Rate, and the Importance of Feature Scaling*, TOWARDS DATA SCI. (July 15, 2020) ("Every time we train a deep learning model, or any neural network for that matter, we're using gradient descent . . . .").

241. *See Loss*, *supra* note 216.

parameters. It does so through the mathematical process known as *gradient descent*.[242]

i. Step 1: Calculate initial loss

Google describes the process of reducing loss as "the 'Hot and Cold' kid's game," where you wander a room while someone shouts "*Warmer!*" Or "*Colder!*" as you get closer or further from a hidden object.[243] In machine learning, "the hidden object is the best possible model."[244] And just like the searcher in the "Hot and Cold" game, the programmer begins "with a wild guess."[245]

Once the programmer has decided on the model's features, he assigns random values to each of the parameters associated with those features and then "wait[s] for the system to tell [him] what the loss is."[246] As explained in more technical detail below, the model then identifies improved parameters, measures their loss, repeats this process until it minimizes loss and the predictions are as accurate as they'll get.[247] Note that in this iterative learning process, the programmer's own predictions—that is, the values of the relevant parameters—are irrelevant.[248] Often, he simply needs to provide the model with an initial guess because for many machine learning problems, "it turns out the starting values aren't important. We could pick random values."[249] From there, the model works its way to a useful prediction on its own.[250]

The first step in this process is to apply a "loss function" that calculates the loss associated with the random parameters selected by the programmer.[251] This will determine how "hot" or "cold" this first iteration of the model is

242. Godoy, *supra* note 240.

243. *Reducing Loss: An Iterative Approach*, GOOGLE FOR DEVELOPERS, https://perma.cc/Y32E-23KJ (archived July 31, 2024).

244. *Id.* (internal quotations omitted).

245. *Id.*

246. *Id.*; *see also Artificial Intelligence and Intellectual Property—Part II: Copyright: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. 74 n.14 (2023) (statement of Mathew Sag) ("[R]andom seeding is important because it helps the model to explore a wide range of possible solutions and to avoid getting stuck in one area of the solution space.").

247. *Reducing Loss: An Iterative Approach*, *supra* note 243.

248. For more on the programmer's role in machine learning, see Lehr & Ohm, note 219 above, at 669-702.

249. *Reducing Loss: An Iterative Approach*, *supra* note 243. Random values suffice for linear regression problems. For nonlinear problems, programmers will have to "babysit" the model to ensure the loss function acts as expected. *CS231n Convolutional Neural Networks for Visual Recognition*, GITHUB, https://perma.cc/RBG7-625X (archived Oct. 20, 2024).

250. *Reducing Loss: An Iterative Approach*, *supra* note 243.

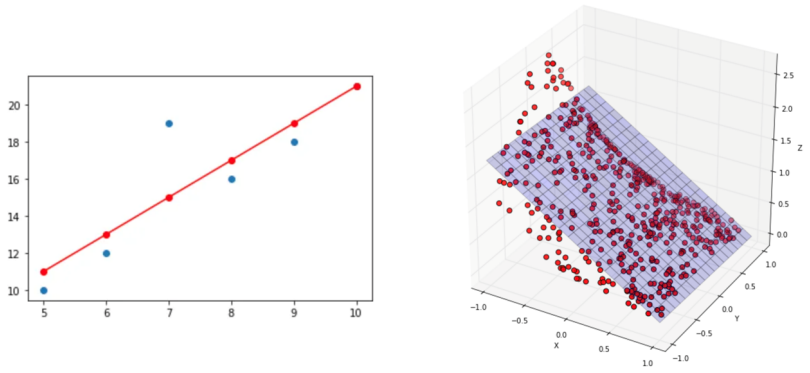
251. *See Loss Function*, *supra* note 216.

compared to the best possible model.[252] To visualize what's happening here, go back to middle school and recall the "line of best fit" you might have drawn on a scatter plot. The loss function is like the scatter plot, only more complex and a little psychedelic.

When a model has only two features and two associated parameters, the line of best fit is operating in two dimensions; this translates to a graph with an x-and y-axis representing the relationship between *x* and *y*. As a result, the line of best fit is a straight line. (Figure 1) But as you add more features, it changes shape. In three dimensions, the line of best fit is "a plane; and in more than three, it's a hyperplane. It's hard to visualize things in hyperspace, but the math works just the same way. In *n* dimensions, we have *n* inputs and the [model] has *n* [parameters]."[253]

Figure 1

Lines-of-Best-Fit in Two Dimensions (Left) and Three Dimensions (Right)[254]



To calculate the loss of the first set of random parameters, the model measures how inaccurate the line of best fit is relative to each of the labeled examples within the batch.[255] Specifically, the loss is measured by the distance between the labeled examples (the individual data points on the graph) and its

252. *Reducing Loss: An Iterative Approach*, *supra* note 243.

253. DOMINGOS, *supra* note 2, at 98-99.

254. Roi Polanitzer, *Data Science One on One—Part 9: Standard Errors of Coefficients*, MEDIUM (Nov. 25, 2021), https://perma.cc/5TTF-4R7F (left); Patrick Wright, *Best-fit Surfaces for 3-Dimensional Data*, INVERSION LABS (Mar. 21, 2016), https://perma.cc/Q45F-G5LP (right).

255. *Loss Function*, *supra* note 216.

prediction based on the provided parameters (the line of best fit).[256] In the left image of Exhibit 1, the line of best fit is the line; in the image on the right, it's the plane. Remember that the line of best fit was drawn based on parameters that the programmer picked at random, so it is likely to be extremely inaccurate at this stage. But that's not a problem because this is just the first step in an iterative learning process that will improve the predictions at each stage.[257] Now having calculated that initial loss, the model can move on to the next step of its training.

ii. Step 2: Gradient descent

In this second step, gradient descent writes the rules. To do so, it uses the loss associated with the random parameters to identify new and improved parameters.[258] These new parameters will reduce the loss and yield more accurate predictions than the random ones the programmer provided.[259] This is the critical step in machine learning because it is here that the model is for the first time generating the values of the parameters independently.[260] This is where machine learning takes humans out of the loop.[261]

At this point, the model functions independently through gradient descent. Gradient descent "iteratively adjusts [the parameters], gradually finding the best combination to minimize loss."[262] To do that, the model must first create a new graph.[263] Instead of plotting the labeled examples and a line of best fit associated with a single set of parameters, as in Figure 1 above, this new graph plots the *value of the loss* associated with a batch of exemplar *parameters*.[264] This is known as aggregate loss.[265]

256. *Backpropagation*, *supra* note 216.

257. *See Reducing Loss: An Iterative Approach*, *supra* note 243.

258. *Id.*

259. *Gradient Descent*, *supra* note 216.

260. *Linear Regression: Gradient Descent*, GOOGLE FOR DEVELOPERS, https://perma.cc/4ECT-49Y2 (last updated Oct. 9, 2024) (explaining how the gradient descent "iteratively finds the [parameters] that produce the model with the lowest loss").

261. This is particularly true because "the training systems are increasingly standardized into a commodity" such that programmers are no longer even writing the algorithms to execute gradient descent. Karpathy, *supra* note 190; *see also Backpropagation*, *supra* note 216.

262. *Gradient Descent*, *supra* note 216.

263. Note that the model is not literally creating a new graph, but implementing sophisticated math. For non-technical readers, the graph is how we can visualize what the model is doing behind the scenes.

264. *Backpropagation*, *supra* note 216.

265. *Award Abstract # 2008532, RI: Small: A Study of New Aggregate Losses for Machine Learning*, U.S. NAT'L SCI. FOUND., https://perma.cc/5MEV-YF5T (archived Oct. 20, 2024).

To calculate aggregate loss, the model adds together the value of the loss associated with each individual labeled example using a given set of parameters.[266] The easiest way to visualize this is when a model has only two features, in which the line of best fit is a line, as we saw in the first step of training. First, the model calculates the value of the loss associated with the line of best fit (i.e., the distance between the points and the line in Figure 2.A).[267] Next, it adds each of those loss values together to get the aggregate loss for that set of parameters.[268] That aggregate loss represents a single point on the new graph (labeled the "starting point" in Figure 2.B).[269] Once the model has calculated the aggregate loss for every possible set of parameters, the new chart will be a U-shaped parabola (Figure 2.B).[270]

266. *Backpropagation*, *Machine Learning Glossary*, *supra* note 216.

267. *Linear Regression: Gradient Descent*, *supra* note 260 (providing a detailed explanation of this process with accompanying charts); *Reducing Loss: Gradient Descent*, GOOGLE FOR DEVELOPERS, https://perma.cc/L8LC-S497 (archived July 31, 2024).

268. *Reducing Loss: Gradient Descent*, *supra* note 267.

269. *Id.*

270. *Id.* This is always true for models with two features. For more complex models, see Part III.B.2.c.iii below.

Figure 2.A
Graph of "Starting Point" Parameters[271]

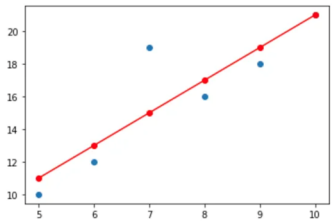
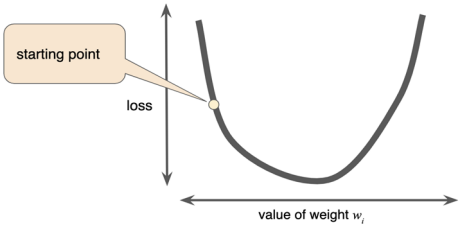


Figure 2.B
Graph of "Aggregate Loss" for All Parameters[272]



Some sets of parameters will undershoot their predictions (on the left side of the U) and some will overshoot them (on the right side of the U), but in two-feature models, there will always be a single set of parameters that minimizes the loss (the bottom of the U).[273] And because improving the accuracy of the prediction means minimizing loss associated with the model, the goal of gradient descent is to find the minimum at the bottom of the U.[274]

To do this, "[t]he first stage in gradient descent is to pick . . . a starting point."[275] As Google explains, "[t]he starting point doesn't matter much; therefore, many algorithms simply . . . pick a random value."[276] From there, the model "calculates the gradient of the loss curve at the starting point" and

271. Polanitzer, *supra* note 254.

272. *Reducing Loss: Gradient Descent*, *supra* note 267.

273. *See Reducing Loss: Learning Rate*, GOOGLE FOR DEVELOPERS, https://perma.cc/C74S-BG7S (archived July 31, 2024).

274. *Reducing Loss: Gradient Descent*, *supra* note 267.

275. *Id.*

276. *Id.*

subsequently "takes a step in the direction of the negative gradient in order to reduce loss as quickly as possible."[277] In other words, if you were to position yourself at any point on the graph, except for the minimum, you would find yourself on a slope from which you could either go uphill or downhill. As may be evident by the name, the goal of gradient *descent* is to go down the hill. That's because we're trying to improve our predictions, and better predictions have lower loss. So the lower you go down the hill the more accurate your predictions get. Thus, with each step the model takes, it identifies new and improved parameters and incorporates those parameters into the model. "The gradient descent then repeats this process, edging ever closer to the minimum."[278] Once it reaches the minimum, the loss function "converges"—the point at which "additional training won't improve the model."[279] The model has identified the parameters associated with the best possible prediction.

In sum, on its journey to convergence, gradient descent has taken the random parameters provided to it by the programmer, and adjusted them on its own, step by step, until it determined parameters that will accurately predict the relationship between the features and the label. From there, the programmer is delivered a working model. The machine has written its own rules.

iii. Gradient descent with complex models

Identifying the lowest possible loss in a model with only two features, such as in the example above, is simple because there will always be a single identifiable minimum.[280] In practice, however, models will often include hundreds, thousands, millions, or billions of features.[281] In such cases, we're no longer operating in two dimensions, but back to operating in psychedelic hyperspace.[282] Although mathematically more complex, gradient descent operates conceptually the same way with complex models, but with a critical distinction for our purposes: In complex models, gradient descent settles on a set of rules that is just one of many generally satisfactory, but mathematically

277. *Id.*

278. *Id.*

279. *Convergence*, *supra* note 216.

280. *Reducing Loss: Gradient Descent*, *supra* note 267 ("Convex problems have only one minimum.").

281. *See, e.g.*, Zhaoxia Deng et al., *Low-Precision Hardware Architectures Meet Recommendation Model Inference at Scale*, IEEE MICRO, Sept.-Oct. 2021, at 93, 93 ("Facebook's production recommendation models consist of many tens to a hundred billion parameters . . . .").

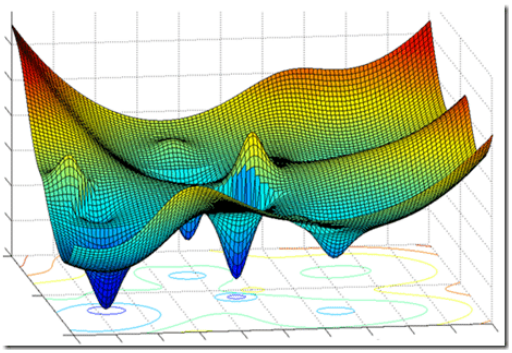
282. DOMINGOS, *supra* note 2, at 98-99.

distinct, sets of rules.[283] Thus, not only does the machine learning programmer outsource the writing of the rules to gradient descent—rules which he cannot explain or understand—but the use of gradient descent for complex models allows him to be agnostic as to the details of those rules.

Though it is hard to visualize complex models with many features, imagine the aggregate loss graph that gradient descent is navigating is not a U, but more like a mountain range.

Figure 3

What the "Aggregate Loss" Landscape Might Look Like in Three Dimensions[284]



In a mountain range, like in Figure 3, just as some peaks are higher than others, some valleys are lower too. Unfortunately, gradient descent is a rather limited sherpa. It only goes down a hill until it finds the bottom.[285] What this means is that if you pick a starting point on the wrong mountain, you might end up at the bottom of a valley that isn't the lowest valley in the mountain range. Instead of the "global minimum"—of which there is only one—you would end up in what programmers call a "local minimum," of which there are many.[286]

In other words, in a model with many features, gradient descent may not result in the lowest possible loss, and a prediction that isn't the most accurate based on the data used to train it. That might lead one to think that allowing a

283. *See id.* at 111.

284. Dheeraj Inampudi & Daniel McKee, Application of Gradient Descent Algorithm in the Program Design Cockpit (PDC) of StratFit 4 fig.1, (unpublished manuscript), https://perma.cc/4MGK-RSW9 (archived Oct. 20, 2024).

285. *Reducing Loss: Gradient Descent*, *supra* note 267 ("The gradient always points in the direction of steepest increase in the loss function.").

286. *See* DOMINGOS, *supra* note 2, at 110-11.

programmer to pick a random starting point wouldn't lead to an accurate prediction, because the odds that you end up in a local minimum instead of a global minimum are extremely high.[287] "But what we've come to realize is that most of the time a local minimum is fine," writes machine learning expert Pedro Domingos.[288] "The error surface [of the graph] often looks like the quills of a porcupine, with many steep peaks and troughs, but it doesn't really matter if we find the absolute lowest trough; any one will do."[289]

Machine learning models that rely on gradient descent therefore have a remarkable tolerance for randomness. From labeled examples and a random set of parameters provided by the programmer, the model generates an aggregate loss graph; and from a random starting point on that graph, it can determine the parameters that are good enough to make useful predictions. The programmer is a necessary part of the process—and requires significant expertise—but he does not play a role in deciding the parameters that define the rules underlying the model's predictions. In fact, with complex models, he can be agnostic as to what parameters are ultimately chosen. Gradient descent has written the rules. And once those rules are written, the model is ready to make predictions.

d. Inference: Making predictions

Once a programmer has trained the model from the labeled examples, the model, resulting in a set of parameters decided upon by gradient descent, is ready to begin making predictions. In machine learning, this process is known as *inference*.[290]

Up until this point, the model has only trained itself on labeled examples. This means that it has defined the mathematical relationships between the features and labels contained within the data fed into it by the programmer.[291] In the inference stage, the model now reviews unlabeled examples, analyzing their features to make predictions about which labels to apply to them.[292] In the spam filter example, the model would analyze each of the features within an email as it comes into your inbox, such as whether the email was sent from an email address in your contacts, the time at which it was sent, and whether it contains words and phrases commonly associated with spam. Based on the

287. *See id.*

288. *Id.* at 111.

289. *Id.*

290. *Inference*, *supra* note 216 ("[T]he process of making predictions by applying a trained model to unlabeled examples.").

291. *See Training*, *supra* note 216 ("During training, a system reads in examples and gradually adjusts parameters.").

292. *Inference*, *supra* note 216.

parameters associated with each of those features, it would predict how likely it is that the email is spam and place it either in your inbox or in the spam folder. For the first time, a piece of unlabeled data has been labeled exclusively by the model.

In the training phase, we saw the first challenge facing the speech status of machine learning output: (1) that the programmer does not write the rules underlying machine learning models—a dramatic departure from how predictions are made with traditional code. The inference phase introduces the second and third challenges: (2) that the programmer cannot explain or, in many cases, even understand how those rules are applied—a conundrum known as the "black box problem"; and (3) that because these rules are probabilistic, they will necessarily be wrong at least some of the time.[293] Thus, machine learning models make predictions based on rules the programmer did not write, which will inevitably and unpredictably produce output contrary to what the programmer intended, for reasons the programmer cannot explain. As we will explore in Part IV, this trio of facts about the output of machine learning models deprives it of speech certainty and places it outside the protection of the First Amendment.

e. Explainability: The "black box" problem

In the inference stage, we finally encounter the output of a machine learning model: its predictions about never-before-seen data—the likelihood that, for example, a new email is spam. Fundamentally, these predictions cannot be explained by the programmer. In most cases, they are "black boxes"—rules so complex that they are incapable of human understanding.[294] In others, they are the product of too large a set of rules for any human to practically

293. Loukides, *supra* note 4 ("[W]e have to be aware that machine learning is never going to be a 100% solution . . . ."); *see also* MONIKA BICKERT, CHARTING A WAY FORWARD: ONLINE CONTENT REGULATION 7 (2020), https://perma.cc/L2PV-S44Y (archived Oct. 20, 2024) ("[I]nternet companies' enforcement of content standards will always be imperfect."); Douek, *supra* note 8, at 764 ("[A] probabilistic conception of online speech acknowledges that enforcement of the rules made as a result of this balancing will never be perfect, and so governance systems should take into account the inevitability of error and choose what kinds of errors to prefer.").

294. Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, ACM COMPUTING SURVS., Aug. 2018, at 1, 5 ("A black box predictor is a data-mining and machine-learning obscure model, whose internals are either unknown to the observer or they are known but uninterpretable by humans." (emphasis omitted)); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PENN. L. REV. 871, 886 (2016) ("[E]ven the original programmers of the algorithm have little idea exactly how or why the generated model creates accurate predictions."); Vikas Hassija et al., *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, 16 COGNITIVE COMPUTATION 45, 53 (2024) ("Most ML models behave as black-box models.").

process.[295] As a result of this inherent opacity in machine learning models, a vibrant field of research has emerged to try to understand how machine learning algorithms work.

Given the interdisciplinary nature of this research and how quickly it is evolving,[296] researchers do not always mean the same thing when they talk about the "explainability" and "interpretability" of a model.[297] To be as clear as possible, what we mean when we say that machine learning predictions cannot be *explained* by the programmer is this: The machine learning programmer cannot explain with the same precision as the traditional programmer why his algorithm produced the prediction that it did. That is, the programmer didn't write the rules, and the complexity of the model that generated the rules means he can't explain them either.

Explainability and interpretability researchers have used the terms "global interpretability,"[298] "global holistic interpretability,"[299] "line of reasoning" explanations,[300] and "decomposability"[301] to describe the ability to explain the

295. Guidotti et al., *supra* note 294, at 9 ("[I]f a too large set of rules, or a too deep and wide tree are returned they could not be humanly manageable even though they are perfectly capturing the internal logic of the black box for the classification.").

296. Hassija et al., *supra* note 294, at 46 ("Methods and techniques have advanced at such a rapid rate that a new field has been created around them: explainable artificial intelligence (XAI).").

297. *See, e.g.*, CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE, ch. 3.4 (2d ed. 2022), https://perma.cc/4KPC-DNEV ("There is no real consensus about what interpretability is in machine learning."); Roberto Confalonieri, Ludovik Coba, Benedikt Wagner & Tarek R. Besold, *A Historical Perspective of Explainable Artificial Intelligence*, WIREs DATA MINING & KNOWLEDGE DISCOVERY, Jan.-Feb. 2021, at 1, 2 ("[T]here is no clear agreement about what an explanation is . . . ."); Zachary C. Lipton, *The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery.*, QUEUE, May-June 2018, at 1, 4 ("[T]he term interpretability is ill-defined . . . ." (emphasis omitted)).

298. Guidotti et al., *supra* note 294, at 6 ("A model may be completely interpretable, i.e., we are able to understand the whole logic of a model and follow the entire reasoning leading to all the different possible outcomes."); Hassija et al., *supra* note 294, at 55 ("Global interpretable approaches are intended to make it easier to comprehend a model's overarching logic as well as the whole justification used to produce specific predictions."). By contrast, "[l]ocal interpretability focuses on providing explanations separately for each choice and prediction rather than providing a detailed description of the intricate mechanism underlying the entire black-box model." Hassija et al., *supra* note 294, at 57.

299. MOLNAR, *supra* note 297, at ch. 3.3.2 ("This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures.").

300. Confalonieri et al., *supra* note 297, at 4 ("Seeing an explanation as a line of reasoning means mainly understanding it as a trace of the way that production or inference rules are used by the system to make a certain decision.").

inner workings of an algorithm in comprehensive detail. The central idea for each of these definitions is "that each part of the model—input, parameter, and calculation—admits an intuitive explanation."[302] They provide "a trace of the way that . . . rules are used by the system to make a certain decision."[303]

In practice, perfect explainability is unachievable in machine learning.[304] In large part, this is because the ability to interpret a machine learning algorithm decreases as its complexity increases.[305] As a result, the trouble that machine learning programmers face in explaining their algorithms is two-fold. First, the "black-box" models that tend to achieve the highest accuracy are simply too complex to be explained.[306] But because the predictions' accuracy is the programmers' primary goal, they often choose to use such models anyway.[307]

Second, even when machine learning programmers use methods of machine learning that are theoretically explainable, they tend not to be explainable in practice.[308] The logic underlying the predictions of such algorithms—in practice, the vast majority of such algorithms—involves so many rules that they cease to be "humanly manageable."[309] And because the machine determined these rules independently,[310] no machine learning programmer (or even team of such programmers) can comprehensively

301. *See* Lipton, *supra* note 297, at 14.

302. *Id.*

303. Confalonieri et al., *supra* note 297, at 4.

304. *See, e.g.*, Levy, *supra* note 15 (quoting Chris Olah, an AI researcher, saying, "we have these systems, we don't know what's going on. It seems crazy.").

305. *See* Guidotti et al., *supra* note 294, at 6 ("[A] component for measuring the interpretability is the complexity of the predictive model in terms of the model size." (emphasis omitted)).

306. Hassija et al., *supra* note 294, at 46 ("AI algorithms suffer from opacity, i.e., the situation in which a system is unable to offer any reason or suitable explanation involved behind its decisions, commonly referred to as 'the black-box problem.'").

307. *See, e.g.*, *infra* note 320 (showing machine learning is widely used across major technology platforms).

308. *Compare* Confalonieri et al., *supra* note 297, at 6 ("[S]ome machine learning models can be considered interpretable by design, namely decision trees, decision rules, and decision tables . . . ."), *with* MOLNAR, *supra* note 297, at ch. 3.3.2 ("Any feature space with more than 3 dimensions is simply inconceivable for humans."), *and* Guidotti et al., *supra* note 294, at 9 (underscoring that models may "not be humanly manageable even though they are perfectly capturing the internal logic of the black box for the classification").

309. Guidotti et al., *supra* note 294, at 9; *see also* Confalonieri et al., *supra* note 297, at 6-7 (noting that "the majority of machine learning models work as black-boxes" that result in "an opaque decision model").

310. *See supra* Part III.B.2.c.

explain their internal logic in the way that a traditional programmer, who wrote his own rules in code, can.[311]

f. The inevitability of errors

The third and final question raised by inference is what to do with the fact that the output of machine learning is guaranteed to be wrong at least some of the time.

Machine learning algorithms are, at their core, probability machines. The labels they apply to never-before-seen data estimate the probability that new data will follow the same patterns as the labeled examples the algorithm was trained upon.[312] But as the saying goes, past results are no guarantee of future performance. And this is as true in machine learning as it is in life. The nature of probability is that we can never be certain about a particular machine learning outcome.[313] We can be highly confident about a model's predictions—perhaps as high as 90%, 99%, or even 99.999%. But in machine learning, the programmer can never predict an outcome with 100% certainty.[314] This fact guarantees the model will make mistakes.

To be sure, programmers can use traditional code to make probabilistic predictions too—predictions that also inevitably get things wrong. But with traditional code, the programmer writes the rules, so even when the predictions are wrong, the algorithm has faithfully executed the rules written by the programmer; the errors can thus fairly be directly attributed to him.[315] With machine learning, the programmer neither writes the algorithm's rules, nor can he explain them.[316] Under such conditions, when the algorithm inevitably makes a mistake—that is, it makes a decision contrary to what the

311. While a less precise explanation may be sufficiently useful for other purposes, only perfect explainability can satisfy the First Amendment's speech certainty principle for programmers who claim algorithmic output as their speech. *See infra* Part IV.

312. *See supra* Part III.B.2.b.

313. MICHAEL J. EVANS & JEFFREY S. ROSENTHAL, PROBABILITY AND STATISTICS: THE SCIENCE OF UNCERTAINTY 1 (2d ed. 2009) ("Probability is the science of uncertainty.").

314. *See* Loukides, *supra* note 4; *see also* Complaint, *supra* note 8, ¶ 35 (showing that X's algorithms could not achieve 100% accuracy); Douek, *supra* note 8, at 764 (recognizing that "governance systems [for online speech] should take into account the inevitability of error").

315. EUGENE VOLOKH & DONALD M. FALK, FIRST AMENDMENT PROTECTION FOR SEARCH ENGINE SEARCH RESULTS 11 (2012) ("These human editorial judgments are responsible for producing the speech . . . .").

316. *See supra* Part III.B.2.e.

programmer intended—one cannot reasonably attribute those mistakes to the programmer.[317]

* * *

Thus, the evolution of code from traditional programming techniques to machine learning models doesn't simply signal a profound change in technical computing; it also marks a distinct difference in how humans relate to the ultimate output of those machines. For the traditional programmer, that output is the direct reflection of the rules they wrote in code. For the machine learning programmer, however, the output depends on rules written by the model. These rules shape the machine learning model's internal logic—logic that machine learning programmers did not write and cannot explain. As a result, when a machine learning model inevitably produces errors—an email mistakenly tagged as spam—those errors cannot be attributed to the machine learning programmer as directly as they can for traditional programmers and the errors of their algorithms. The traditional programmer's authorship of the probabilistic code means he can know with certainty that the output reflects the rules he wrote. Its errors are his errors.[318] The machine learning programmer cannot say the same. The output of his model will make errors that he did not intend or anticipate, emerging from rules he did not write and cannot explain. As will be explored in Part IV below, these features pose serious challenges to his efforts to claim this output as his speech.

C. Platforms' Purported Machine Learning "Speech"

The machine learning process described above increasingly underpins much of our digital experience online, from social media to entertainment streaming platforms to the emerging use of generative artificial intelligence.[319] While we believe that the principles in this paper apply broadly, a comprehensive survey of machine learning's applications is beyond the scope of this paper. Instead, this Subpart briefly spotlights three "paradigmatic" functions of machine learning on social-media platforms: the ranking,

317. Lessig, *supra* note 23, at 276 ("At some point along the continuum between your first program, 'Hello world!,' and [artificial intelligence], the speech of machines crosses over from speech properly attributable to the coders to speech no longer attributable to the coders.").

318. Similarly, his authorship means that any errors arising from bugs in the code can be directly attributed to him. He may not have *meant* for certain errors in the output to arise, but they nonetheless arose *because* of errors in his code.

319. Meta Careers, *Machine Learning at Meta*, META (Oct. 13, 2022), https://perma.cc/R4DK-QVD4; *Machine Learning: Learning How to Entertain the World*, NETFLIX RSCH., https://perma.cc/RHV4-DZM6 (archived Oct. 20, 2024); Goodrow, *supra* note 10 (discussing machine learning as it is deployed on YouTube); *How Should AI Systems Behave, and Who Should Decide?*, OPENAI (Feb. 16, 2023), https://perma.cc/6RTN-CEM2.

recommendation, and removal of content.[320] These three applications are some of the means by which these platforms perform content moderation.[321] Consequently, they also capture most of the activities that these platforms claim as their speech.[322]

1. Ranking

Platforms use ranking models to determine the order in which content is shown to a user.[323] Essentially, ranking models decide which of the multitudes of content you *could* see, you actually do see, and in what order.[324] For social media platforms, models are generally trained on posts a user has previously interacted with, analyzing the features of the content alongside user engagement data to determine the likelihood that a given user will find a new piece of content interesting or useful.[325] The model analyzes a set of content

320. *See, e.g.*, Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2393-94 (2024); Singh, *supra* note 6 ("[M]any [platforms] have developed or adopted automated tools to enhance their content moderation practices, many of which are fueled by artificial intelligence and machine learning."); DOMINGOS, *supra* note 2, at 152 ("Google uses machine learning in every nook and cranny of what it does."); Pandu Nayak, *How AI Powers Great Search Results*, GOOGLE KEYWORD (Feb. 3, 2022), https://perma.cc/73NU-KV65 ("Thanks to advancements in AI and machine learning, our Search systems are understanding human language better than ever before."); Will Knight, *Facebook's Head of AI Says the Field Will Soon 'Hit the Wall'*, WIRED (Dec. 4, 2019, 7:00 AM), https://perma.cc/84BC-H5DV ("The two core uses of AI today in Facebook are making the platform safer for users and making sure what we show users is valuable to them."); Yi Zhuang, Arvind Thiagarajan & Tim Sweeney, *Ranking Tweets with TensorFlow*, TENSORFLOW BLOG (Mar. 4, 2019), https://perma.cc/BNN2-A6LH (describing "the machine learning system we use to rank Twitter's home timeline").

321. *See* Gillespie, *supra* note 6, at 1.

322. *See, e.g.*, Brief for Respondents, *supra* note 1, at 23 (arguing that social media platforms "engage[] in speech when disseminating 'curated compilations of speech' created by others" (quoting NetChoice, LLC v. Att'y Gen., 34 F.4th 1196, 1213 (11th Cir. 2022))); Brief for Petitioners, *supra* note 1, at 27 (arguing that social media platforms' editorial decisions are protected even when executed via algorithm).

323. *See* Francesco Casalegno, *Learning to Rank: A Complete Guide to Ranking Using Machine Learning*, MEDIUM (Feb. 28, 2022), https://perma.cc/7Z6G-RQL9; *Ranking*, *supra* note 216 (defining ranking as "[a] type of supervised learning whose objective is to order a list of items").

324. *Ranking*, *supra* note 216.

325. *See, e.g.*, *Our Approach to Facebook Feed Ranking*, META, https://perma.cc/H3XA-3SP2 (archived Oct. 20, 2024); Dunn, *supra* note 207 ("Machine learning models are part of ranking and personalizing News Feed stories, filtering out offensive content, highlighting trending topics, ranking search results, and much more."); *Twitter's Recommendation Algorithm*, *supra* note 196 ("Ranking is achieved with a ~48M parameter neural network that is continuously trained on Tweet interactions to optimize for positive engagement (e.g., Likes, Retweets, and Replies). This ranking mechanism takes into account thousands of features and outputs ten labels to give each

footnote continued on next page

and then presents it in a certain order for the user based on a given set of criteria.[326] All the major platforms that present a feed of content—including Facebook, X/Twitter, and Google—use machine learning to rank the content they show you.[327]

2. Recommendation

Social media platforms also rely on machine learning models to recommend content to their users.[328] This is the "secret sauce" that powers which TikTok or YouTube videos are shown to you,[329] determines which Netflix shows and movies are suggested to you,[330] and underpins why you see content from accounts you don't follow on Facebook or X.[331] These models introduce content to you that you did not expressly seek out yourself.[332] Although the technical details of recommendation models differ from ranking models, they function similarly: Both models analyze the features of the content you've engaged with and how you engaged with it to determine the likelihood that you will find a different piece of content engaging.[333]

3. Removal

Social media platforms rely on machine learning models to enforce their content policies by identifying and removing offending content. At Facebook, "[f]or example, an AI model predicts whether a piece of content is hate speech

Tweet a score, where each label represents the probability of an engagement. We rank the Tweets from these scores.").

326. *See, e.g.*, *Twitter's Recommendation Algorithm, supra* note 196.

327. *See, e.g.*, *id.*; *Our Approach to Facebook Feed Ranking*, *supra* note 325; *Automatically Generating and Ranking Results*, GOOGLE SEARCH, https://perma.cc/ZM46-XWNQ (archived Oct. 20, 2024).

328. Chris Meserole, *How Do Recommender Systems Work on Digital Platforms?*, BROOKINGS (Sept. 21, 2022), https://perma.cc/KR7C-G9EV.

329. *See* Narayanan, *supra* note 196; Goodrow, *supra* note 10.

330. *See* Christopher Mims, *How Netflix's Algorithms and Tech Feed Its Success,* WALL ST. J. (July 28, 2023), https://perma.cc/4632-DVAS.

331. *Our Approach to Facebook Feed Ranking*, *supra* note 325; *Twitter's Recommendation Algorithm*, *supra* note 196.

332. *See, e.g.*, *Recommendation System*, *supra* note 216 (listing "[m]ovies that similar users have rated or watched" as an example of a recommendation system).

333. Note that in other contexts the recommendations may be based on other criteria. For an app store, for example, a recommendation model may recommend other apps based on their similarity to the app you're looking at rather than based on your user data. *See, e.g.*, *Recommendations: What and Why?*, GOOGLE FOR DEVELOPERS, https://perma.cc/JGH2-Z53Y (archived Oct. 20, 2024).

or violent and graphic content."[334] The same is true at YouTube, where "models are trained to identify potentially violative content."[335] Having identified violative content, the models may remove content directly or flag it for human review, depending on the platform's philosophy on content moderation and its capacity to employ human moderators.[336] The removal process generally occurs after the content has been published by a user, but certain content may be removed before the content has been distributed to other users.[337]

IV. Because Machine Learning "Speech" Lacks Speech Certainty, It Is Not Protected By the First Amendment

A. Speech Certainty Is a Threshold Question to the Protection Analysis

Notable scholars have argued that algorithmic output is protected by the First Amendment under both speakers' rights and listeners' rights frameworks.[338] These arguments presume that this output is "speech" entitled to such protection.[339] But, as we illustrate below, that presumption is no longer sound. Before we ask whether something is protected by the First Amendment, we must first confirm that it is "speech" entitled to that protection. This "speech" inquiry—whether something qualifies as "speech" within the meaning of the First Amendment—is an antecedent to the "protection" inquiry. Indeed,

334. *How Enforcement Technology Works*, META, https://perma.cc/3KBM-SUCU (archived Oct. 20, 2024).

335. Matt Halprin & Jennifer Flannery O'Connor, *On Policy Development at YouTube*, INSIDE YOUTUBE (Dec. 1, 2022), https://perma.cc/AR2R-JYM8.

336. *See, e.g.*, *How Automation Is Used in Content Moderation*, GOOGLE, https://perma.cc/X2BB-3GFN ("The policy-violating content is either removed by Google's AI or, where a more nuanced determination is required, it is flagged for further review by trained operators and analysts . . . ."); *cf.* Katie Paul & Sheila Dang, *Exclusive: Twitter Leans on Automation to Moderate Content as Harmful Speech Surges*, REUTERS (Dec. 5, 2022, 1:41 PM PST), https://perma.cc/XWW6-NQKZ (describing X/Twitter's increasing reliance on machine learning to conduct content moderation).

337. *See, e.g.*, *The Four Rs of Responsibility, Part 1: Removing Harmful Content*, INSIDE YOUTUBE (Sept. 3, 2019), https://perma.cc/VG2Z-NJ68 ("We go to great lengths to make sure content that breaks our rules isn't widely viewed, or even viewed at all, before it's removed.").

338. VOLOKH & FALK, *supra* note 315, at 3-6 (speakers' rights); Eugene Volokh, Mark A. Lemley & Peter Henderson, *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651, 654 (2023) (listeners' rights); Benjamin, *supra* note 112, at 1447 (speakers' rights).

339. *See, e.g.*, Volokh et al., *supra* note 338, at 654 (noting that "the right to receive speech" justifies protecting AI output).

First Amendment rights only attach if there's any speech to protect in the first place.[340]

The speech certainty principle provides the tools to address the "speech" inquiry. Once purported speech satisfies the principle, the principle's job is done, leaving the "protection" inquiry to other tools of First Amendment analysis. Speech certainty is therefore entirely compatible with the full spectrum of speakers' and listeners' rights frameworks for First Amendment protection—it simply precedes them in the analysis.[341] Under the speakers' rights framework, a speaker must know what she says when she says it; if she does, we can then ask what strand of First Amendment doctrine protects her right to say it. Under the listeners' rights framework, the threshold question is whether what the listener claims she has a right to receive is speech.[342] If a listener claims a First Amendment right to receive purported speech, that right only attaches if some speaker knew what she said when she said it—and it could subsequently reach the listener.[343]

---

340. *See* 303 Creative LLC v. Elenis, 143 S. Ct. 2298, 2312 (2023) ("All manner of *speech*—from 'pictures, films, paintings, drawings, and engravings,' to 'oral utterance and the printed word'—qualify for the First Amendment's protections." (emphasis added) (quoting Kaplan v. California, 413 U.S. 115, 119-20 (1973))).

341. *See, e.g.*, Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1926 (2019) (assessing whether a public access channel is a government actor as a prior step to First Amendment analysis); Knight First Amend. Inst. at Columbia Univ. v. Trump, 928 F.3d 226, 237 (2d Cir. 2019), *cert. granted*, *vacated sub nom.* Biden v. Knight First Amend. Inst. at Columbia Univ., 141 S. Ct. 1220 (2021) (applying forum doctrine to the social media accounts of government officials); Susan P. Crawford, *First Amendment Common Sense,* 127 HARV. L. REV. 2343, 2346 (2014) (advocating for the application of common carrier doctrine to high-speed internet access providers).

342. Va. State Bd. of Pharmacy v. Va. Citizens Consumer Council, Inc., 425 U.S. 748, 756 (1976) (concluding that commercial speech is protected because "protection afforded is to the communication, to its source and to its recipients both"); *see also* Animal Legal Def. Fund v. Kelly, 434 F. Supp. 3d 974, 995 (D. Kan. 2020) ("The right to receive information is entirely derivative of—and cannot enlarge—the willing speaker's rights.").

343. Several courts have discussed listeners' rights under the First Amendment. *See Va. Citizens Consumer Council, Inc.*, 425 U.S. at 756 (concluding that commercial speech is protected because "protection afforded is to the communication, to its source and to its recipients both"); Lamont v. Postmaster Gen., 381 U.S. 301, 305, 307 (1965) (relying on "the addressee's First Amendment rights" to receive mail, rather than the sender's right to send it, where the sender was a foreign government); Citizens United v. FEC, 558 U.S. 310, 341 (2010) ("The First Amendment protects speech and speaker, and the ideas that flow from each."); *id.* at 473 (Stevens, J., concurring in part and dissenting in part) ("But when the speakers in question are not real people and when the appeal to 'First Amendment principles' depends almost entirely on the listeners' perspective, it becomes necessary to consider how listeners will actually be affected." (citation omitted)); *Lamont*, 381 U.S. at 307 ("We rest on the narrow ground that the addressee in order to receive his mail must request in writing that it be delivered.").

As a vivid example, imagine that the government planned to raze a forest to build a highway and an environmental protection group raised a listeners' rights claim to enjoin the plan by claiming it would prevent them from hearing the wind rustling through the trees. Of course, their argument would fail because, for a multitude of reasons, the wind rustling through the trees is not speech.[344] Although a listener may hear something, and may even impute a message to it, she does not have a First Amendment right to hear it unless it is "speech" within the meaning of the First Amendment.[345]

As we'll show in this section, the output of machine learning models, like the wind rustling through the trees (though admittedly a much closer question), is not speech.[346] The purported speaker—the programmer—cannot know with certainty what the output is when it is generated, so she cannot claim it as her speech. Nor can anybody else.[347] Because it lacks speech certainty, it is simply not speech within the meaning of the First Amendment. By extension, the fact that machine learning output is not speech prevents anyone—speaker or listener—from invoking First Amendment protection for it.

B. Assessing the Speech Certainty and Protection of Algorithmic Output

In lawsuits across the country, social media platforms have claimed that the output of their algorithms is protected under the First Amendment.[348] To

344. One reason, for example, is because no speaker has articulated that speech. *See* LARRY ALEXANDER, IS THERE A RIGHT OF FREEDOM OF EXPRESSION? 8-9 (2005) (recognizing that a sunset has no speaker). For an alternative view that the First Amendment protects listeners' rights regardless of the existence of a speaker, see *id.* (recognizing that restrictions on viewing a sunset could arguably implicate freedom of expression).

345. *See Va. Citizens Consumer Council*, 425 U.S. at 756 (concluding that commercial speech is protected because "where a speaker exists . . . protection afforded is to the communication, to its source and to its recipients both"); *see also Animal Legal Defense Fund*, 434 F. Supp. 3d at 995 ("The right to receive information is entirely derivative of—and cannot enlarge—the willing speaker's rights.").

346. As discussed in Part V, recognizing the non-speech status of something means that the government is free to regulate it. If and how the government should do so are separate questions that certainly raise more challenging and complex questions in the context of regulating machine learning algorithms than in the context of, say, regulating the wind.

347. The only other candidate to claim it would be the model itself. To date, we have not granted autonomous algorithms themselves their speech rights. *See, e.g.*, LESSIG, *supra* note 23, at 280 ("Talking cats have no First Amendment rights, as the Eleventh Circuit has informed us. Likewise, . . . if we see replicants as a kind of animal, then their speech, too, should be entitled to no strong First Amendment protection."); *cf.* Thaler v. Perlmutter, 687 F. Supp. 3d 140, 146 (D.D.C. 2023) (recognizing human authorship as a "bedrock requirement of copyright").

348. *See, e.g.*, NetChoice, L.L.C. v. Paxton, 49 F.4th 439, 464 (5th Cir. 2022); NetChoice, LLC v. Att'y Gen., 34 F.4th 1196, 1203 (11th Cir. 2022).

make this claim, platforms have primarily relied on two legal doctrines: editorial discretion and expressive conduct.[349] Critically, however, the plaintiffs, the platforms, and the Courts have not addressed the fact that the way these algorithms operate has shifted from traditional code to machine learning over the past decade.

As we explore below, a recognition of this shift dramatically alters the analysis. To illustrate the difference, we explore the protection of speech of hypothetical algorithms created through three distinct methods: traditional code, traditional code used to make predictions, and a machine learning model. Each hypothetical focuses on a social media platform that allows users to post content, employs an algorithm to enforce its content guidelines, and then publishes posts in the platform's feed.[350]

For each hypothetical, we first ask the threshold question of whether the algorithmic output is characterized by speech certainty. That is, does the programmer know with certainty what the output will be at the moment it is produced? The answer to this question determines whether or not the "speech" is speech for purposes of the First Amendment. Next, we ask whether the purported speech is *protected* speech under the doctrines of editorial discretion and expressive conduct. For editorial discretion, the decisive question is whether the platform's use of an algorithm qualifies as an exercise of the platform's judgment as to the contents of the compilation.[351] Specifically, does the algorithm guarantee that the platform will not publish anything in the feed

349. *NetChoice*, 49 F.4th at 451, 464; *NetChoice*, 34 F.4th at 1210. Because platforms have not relied on listener's rights, this Subpart only explicitly analyzes the algorithms against these two speaker's rights frameworks. As addressed in Part IV.A, however, because we conclude that the output of machine learning lacks speech certainty, it cannot satisfy the "speech" inquiry that is an antecedent to the "protection" inquiry. Thus, a listener's rights claim would fail because the listener wouldn't have any "speech" to claim a right to receive.

350. Importantly, we distinguish between the initial output of a machine learning algorithm such as posts on a user's timeline, and the post-hoc removal of content or deplatforming that humans execute after the algorithm has made its initial determination on the content through its model. Such decisions—those dictated directly by humans—are unquestionably characterized by speech certainty.

351. In editorial discretion terms, each platform employs an algorithm to determine whether to include or exclude content from a published compilation of its users' speech. In doing so, they all easily satisfy two of the three criteria needed for First Amendment protection under the doctrine of editorial discretion: (1) the platforms' feeds of non-violating content are compilations of speech and (2) those compilations are published when they are served to a user. *See* Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2402 (2024). Thus, the narrow question we focus on for each hypothetical is the thornier third prong of the editorial discretion analysis: whether the platform's use of an algorithm qualifies as an exercise of the platform's judgment as to the contents of the compilation. *Id.*

"which their 'reason' tells them should not be published"?[352] And for expressive conduct, we subject the algorithm to the *Spence* test to determine whether (1) the speaker intended to convey a particularized message via the algorithm and (2) there is a great likelihood that the message would be understood by those who viewed its output.

We conclude that unlike algorithms written with traditional code—which consensus rightly views as protected speech[353]—machine learning algorithms lack speech certainty and cannot earn protection under the doctrines of editorial discretion or expressive conduct. This is consistent with the unspoken (and until now unnecessary) understanding that speech protected by the First Amendment must be characterized by speech certainty.

1. Traditional code

Our first hypothetical shows that a platform's use of an algorithm written with traditional code is consistent with the First Amendment's underlying principle of speech certainty and protected by the doctrines of editorial discretion and expressive conduct.

Imagine a programmer who hates vegetables has developed a social network for like-minded carnivores called MeatUp. In her content guidelines, she decides that the word "eggplant" is not allowed on MeatUp and writes code to reflect the following rule to govern the platform:

> If a user attempts to publish a post containing the word "eggplant," serve the user with an error message reading, 'Sorry, this is an eggplant-free zone.' Otherwise, publish the post to the MeatUp feed.

a. The output of traditional code is characterized by speech certainty

By writing traditional code to automate her publication decisions, the programmer can guarantee that the code will faithfully execute the rule she wrote to determine a post's fitness for publication—namely, that no posts containing the word "eggplant" will be included in the MeatUp feed.[354] This

352. Associated Press v. United States, 326 U.S. 1, 20 n.18 (1945); Mia. Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 256 (1974) ("Compelling editors or publishers to publish that which reason tells them should not be published is what is at issue in this case." (internal quotation marks omitted)).

353. *See, e.g.*, VOLOKH & FALK, *supra* note 315, at 12-13; Benjamin, *supra* note 112, at 1447. Although the speech status of the output of traditional code is not in dispute, we walk through it to show that it is consistent with the principle of speech certainty *and* to articulate the logic underlying its protection. Only by holding machine learning algorithms against traditional code can one see why the logic that protects the latter does not extend to the former.

354. *See supra* Part III.B.1.

guarantee means that at all times, the MeatUp programmer knows what her speech—the collection of posts in the MeatUp feed—will be at the moment it is published.[355] She knows that it will always, without fail, be a compilation of speech that excludes any mention of the word "eggplant." The output of her traditional code is therefore characterized by speech certainty, bringing it within the scope of the First Amendment's protection.

b. The output of traditional code is protected editorial discretion

The consensus view correctly holds that the doctrine of editorial discretion protects the output of the MeatUp programmer's code as the programmer's speech.[356] In editorial discretion terms, the MeatUp code is incapable of doing anything other than including what the programmer intended to include (all posts without the word "eggplant") and excluding what she intended to exclude (all posts with the word "eggplant").[357] The use of traditional code merely automates the same editorial decisions that the programmer would have made herself.[358] And as Eugene Volokh and Donald M. Falk explain, "[s]uch automation does not reduce the First Amendment protection."[359] Indeed, the use of code is a direct exercise of the programmer's "right as a private speaker to shape her expression by speaking on one subject while remaining silent on another."[360] It guarantees that every time the programmer's algorithm publishes the MeatUp feed, the feed will perfectly reflect her editorial judgments as to its contents.

c. The output of traditional code is protected expressive conduct

The output of the traditional code is also protected as expressive conduct under *Spence*'s two-pronged test. On the first prong, the programmer's intent to convey a particularized message (that the platform doesn't endorse eggplant-

355. It's true that because the programmer doesn't review every user post prior to its publication in the MeatUp feed, she doesn't know with any specificity the content of users' posts that make up the platform's speech. But the consensus view is that this ultimately doesn't matter for First Amendment purposes because her "human editorial judgments" expressed in the code "are responsible for producing the [platform's] speech"—the compilation of user posts she decides to publish. *See* VOLOKH & FALK, *supra* note 315, at 11.

356. *See, e.g.*, *id.*; Benjamin, *supra* note 112, at 1467.

357. *See supra* Part II.A.

358. *See* Columbia Broad. Sys., Inc. v. Democratic Nat'l Comm., 412 U.S. 94, 124 (1973) ("[E]diting is selection and choice of material.").

359. VOLOKH & FALK, *supra* note 315, at 11.

360. Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 574 (1995) (explaining that First Amendment protection extends to private groups excluding unwanted messages from parades).

based content) is effectively conveyed through the conduct (the algorithm's exclusion of the word "eggplant") because the programmer can be certain of the conduct.[361] The intended message is intertwined with the intentional conduct. That is, the programmer wrote code with the goal of excluding the word "eggplant" from her platform and the code can do nothing but faithfully accomplish that goal, thereby communicating the message via the expressive conduct.[362]

Depending on the context, the second prong of *Spence* is also likely satisfied. A reasonable observer would be able to tell that the programmer disapproves of eggplant-content when her eggplant-laden posts are consistently rejected for publication on the platform.[363] Even if the reasonable observer isn't the one posting eggplant-content, she might still be able to understand the MeatUp programmer's anti-eggplant message by viewing the platform and noting the total absence of the word 'eggplant.'[364] Thus, even though the conduct results in the *absence* of content from the ultimate compilation, both prongs of the *Spence* test are likely satisfied.

2. Probabilistic traditional code

The same rationale that protects the output of a simple algorithm exists to protect the output of more complex algorithms created with traditional code, such as those that calculate probabilities to make predictions ("probabilistic traditional code").[365] To illustrate why, let's introduce our second hypothetical: a social media platform for stone fruit enthusiasts that only allows users to post images of plums called PlumsUp. To enforce this plums-only rule, the PlumsUp programmer determines through her own statistical analysis that an image is 78% likely to be a plum when at least 70% of the image is purple.[366] And because she wants her platform to only include images of plums, she writes an algorithm in code that reflects the following rule:

> If a user attempts to publish a photo, analyze the colors of the image, and if at least 70% of the image is purple, publish it. Otherwise, serve the user with an error message reading, "Sorry, this is a plums-only zone."

361. *See supra* Part II.B.1 (explaining the first prong of the *Spence* test).

362. *See supra* Part II.B.1.

363. *See supra* Part II.B.2 (explaining the second prong of the *Spence* test).

364. *See supra* Part II.B.2.

365. *See supra* Part III.B.1.

366. Whether or not this is a good rule is irrelevant. Mia. Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 258 (1974) ("The choice of material to go into a newspaper, and the decisions made as to limitations on the size and content of the paper, and treatment of public issues and public officials—*whether fair or unfair*—constitute the exercise of editorial control and judgment." (emphasis added)).

a. The output of probabilistic traditional code is characterized by speech certainty

The speech certainty analysis for PlumsUp's probabilistic traditional code is identical to that of MeatUp's traditional code. Although the algorithm involves statistical analysis to generate probabilities that determine the output—meaning the output will be wrong at least some of the time—the fact remains that the programmer has written traditional code to automate her publication decisions.[367] As a result, she can guarantee that the code will faithfully execute the rule that she wrote, however complex, to determine a post's fitness for publication.[368] This guarantee means that at all times, the PlumsUp programmer knows what her speech will be at the moment it is published; it will be a compilation of posts in which at least 70% of the image is purple. And because all that the PlumsUp programmer can claim as her speech is that compilation—not the individual posts posted by PlumsUp users—it doesn't matter that she can't know specifically what an image depicts at the moment it's published.[369] Because she wrote the algorithm with traditional code, she can be certain that when the algorithm publishes any image, it did so because the image complied with the rules she wrote that shape her speech. The output of her probabilistic traditional code is therefore characterized by speech certainty, bringing it within the scope of the First Amendment's protection.

Moreover, the fact that probabilistic traditional code will, to some extent, make mistakes does not undermine the speech certainty inherent to it.[370] Perhaps, for example, a photo of eggplants or lilacs is published due to its dominant purple tones. What matters for speech certainty purposes is not whether the algorithm accurately predicts plums or not, but whether the algorithm executes the programmer's code as written.[371] Because traditional programming can do nothing but execute the code as written, and because the PlumsUp programmer wrote the code, she can be certain that the output will yield only those posts which she told it to yield.[372] She can be certain about the contents of his speech at the moment it is produced.

---

367. *See supra* Part III.B.1.

368. *See supra* Part III.B.1.

369. *See supra* note 116 (explaining that the doctrine of editorial discretion protects compilations of speech); *see also supra* note 140 (distinguishing the protection of a compilation and liability for the contents of that compilation).

370. *Supra* Part III.B.2.f.

371. Note that this means that even unexpected bugs in the code don't undermine the speech certainty because the machine is merely executing the code as written by the programmer.

372. *Cf.* VOLOKH & FALK, *supra* note 315, at 11 ("These human editorial judgments are responsible for producing the speech . . . .").

b. The output of probabilistic traditional code is protected editorial discretion

The use of probabilistic traditional code raises a new challenge for the editorial discretion analysis, however. The programmer intends for her social media platform to be a plums-only zone, but employs an algorithm that predicts whether an image contains a plum with only 78% accuracy. This means that 22% of the time, the algorithm will publish an image of something other than a plum—an eggplant, perhaps—merely because 70% or more of the image is purple. The programmer says the platform should not publish anything that is not a plum ("this is a plums-only zone"), but it nonetheless publishes eggplants.

If the doctrine of editorial discretion extends only to speakers' published compilations that exclude "anything which their 'reason' tells them should not be published,"[373] does it protect the output of the PlumsUp algorithm?

The consensus answer, which we agree with, is yes: The output of probabilistic traditional code is protected editorial discretion.[374] As in all editorial discretion cases, the relevant unit of speech is the published compilation, which for platforms is the *output* of their algorithms.[375] The PlumsUp programmer, for example, may have begun the process of writing an algorithm with the goal of allowing nothing but plums on her platform, but that *goal* cannot be said to be her speech. Instead, her chosen means for deciding what to include or exclude for publication was an algorithm that unfailingly followed the precise rules she wrote for it in code. Those rules reflect her judgment. And that judgment resulted in the published compilation of photos of mostly, but not exclusively, plums. That compilation, not the goals articulated in her content guidelines, is what she can claim as her speech.[376] And because the programmer used traditional code to write these rules, she can be certain that the compilation of images in the feed will perfectly reflect her editorial judgments as to its contents. It will include what she intended to include (images that are at least 70% purple) and exclude the rest.

Thus, the doctrine of editorial discretion protects the output of traditional code, even as it veers into probabilistic predictions in which the algorithm's output does not perfectly reflect the programmer's goals. The programmer

373. Associated Press v. United States, 326 U.S. 1, 20 n.18 (1945).

374. *See, e.g.*, VOLOKH & FALK, *supra* note 315, at 11; Benjamin, *supra* note 112, at 1466-67.

375. *See supra* note 116 and accompanying text.

376. *See* VOLOKH & FALK, *supra* note 315, at 11 ("These human editorial judgments are responsible for producing the speech . . . ."); *see also* Herbert v. Lando, 441 U.S. 153, 178 (1979) (Powell, J., concurring) ("[W]hatever protection the 'exercise of editorial judgment' enjoys depends entirely on the protection the First Amendment accords the product of this judgment, namely, published speech.").

wrote the algorithm with traditional code, and it can do no more or less than execute the programmer's judgment as expressed in that code.

c. Probabilistic traditional code may qualify as expressive conduct

Whether probabilistic traditional code passes muster under *Spence* is less clear. The first prong is relatively straightforward. The programmer of probabilistic traditional code can be certain of her expressive conduct—that is, that the algorithm will execute the rules precisely as she wrote them.[377] And because she is certain, when she writes traditional code to enforce a rule, she has the requisite intent to communicate a particularized message reflecting that rule.[378] The PlumsUp programmer, for example, intended to create a plums-only zone and wrote a rule using probabilistic traditional code to deliver that vision to her satisfaction. The operation of the PlumsUp algorithm—which yields mostly, but not exclusively plums—is conduct that reflects the programmers' intent to communicate a particularized pro-stone fruit message.[379]

The second prong, however, introduces tougher questions. If the output of probabilistic traditional code necessarily contains mistakes at least some of the time, it calls into question whether a reasonable observer could understand the programmer's message[380]—a subjective line-drawing exercise by any measure. If the algorithm publishes photos of lilacs and eggplants to the platform's feed, then the pro-stone fruit message might be lost on an end user. The question then becomes one of degree—how many lilacs and eggplants are sufficient to dilute the message such that the conduct is no longer expressive?

While the speech certainty of the PlumsUp algorithm's output makes it speech for purposes of the First Amendment, it's unclear whether that speech could be properly characterized as protected expressive conduct under current doctrine. This doctrinal gray area isn't ultimately material, however, because the output of probabilistic traditional code is plainly protected under the doctrine of editorial discretion.[381] Nonetheless, it exposes for the first time the shaky First Amendment grounds on which probabilistic speech stands.

377. *Supra* Part III.B.1.

378. *See supra* Part II.B.1 (explaining the first prong of the *Spence* test).

379. *See supra* Part II.B.1.

380. *See supra* Part II.B.2 (explaining the second prong of the *Spence* test).

381. *Supra* Part IV.B.2.b.

3. Machine learning

The logic that protects the output of traditional code under the doctrines of editorial discretion and expressive conduct falters when applied to the output of machine learning models. As this third and final hypothetical will illustrate, platforms that rely on machine learning models cannot claim their output to be protected under the doctrines of editorial discretion or expressive conduct. Indeed, because this output is not characterized by speech certainty, it is not even speech for purposes of the First Amendment.[382]

In this hypothetical, imagine a programmer who decides to compete directly with the PlumsUp social network and launches her own plums-only platform. But rather than using traditional code to determine whether a photo contains a plum (and therefore is fit for publication), she opts to use a machine learning model. She calls her platform PlumGPT.

Instead of writing a set of instructions in traditional code to express a set of rules determined by the programmer, as the MeatUp and PlumsUp programmers did, the PlumGPT programmer first trains a model to determine those rules for her.[383] By the end of the training process, the PlumGPT model can identify whether a photo contains a plum with an 88% success rate—a 10% improvement over the PlumsUp algorithm. Then she writes an algorithm in *traditional* code that incorporates the PlumGPT model to enforce those rules on the PlumGPT platform:

> If a user attempts to publish a photo, analyze the photo with the PlumGPT model. If the PlumGPT model determines the image is a plum, publish it. Otherwise, serve the user with an error message reading, "This is a plums-only zone."

a. Machine learning output is not characterized by speech certainty

Once again, the purported speech is the output of the algorithm. But unlike traditional programmers, the machine learning programmer cannot know with certainty what the output of her algorithm will be at the moment it is generated.

First, the machine learning programmer's role in developing her algorithm lacks the direct connection to the algorithm's output as traditional programmers and their algorithms.[384] As explained above, it is that direct connection—that the algorithm executes the rules as written by the

382. *Supra* Part I (explaining why the text, history, and purposes of the First Amendment compel recognition of the speech certainty principle); *supra* Part II (explaining why relevant First Amendment precedent does the same).

383. *See supra* Parts IV.B.1-.2.

384. *See generally supra* Part III.B.2 (distinguishing machine learning from traditional programming).

programmer—that allows us to call the output of traditional algorithms the programmer's speech.[385] This is particularly true for probabilistic traditional algorithms in which the output is guaranteed to be wrong at least some of the time.[386] When probabilistic traditional algorithms make mistakes, its programmer can be certain about why they occurred; namely, because it followed the instructions given to it by the programmer.[387] When a machine learning model makes mistakes, however, its programmer cannot say the same. This is because she neither wrote the rules that determined the output,[388] nor can she fully explain them;[389] the machine learning programmer intended for the model to only say X, but the rules written by the model (independently of the programmer) led it to say Y.

Here, as explained in Part III and briefly above, the PlumGPT programmer did not write the rules that determine whether a photo contains a plum. She trained the machine learning model through a great deal of hard work that required a great deal of expertise. But she did not at any point define for the model how it should predict whether a photo contains a plum or not. That task—defining the rules underlying the model's predictions—was executed by gradient descent when it identified the set of parameters that defined the relationships between each of the features in the model.[390] Thus, while the PlumGPT programmer has some level of influence over the output, she lacks the direct connection that ensures she can be certain of its output.

Moreover, the machine learning programmer is also incapable of comprehensively explaining or understanding the rules underlying the model in the way that a traditional programmer can. In machine learning, because explainability can never be perfect, it is discussed in terms of a spectrum—measured in terms of *how* explainable a model is.[391] For speech certainty purposes, however, it is a binary: The PlumGPT programmer can either explain her algorithm fully and its output is her speech, or she can't and it isn't.[392]

---

385. *Supra* Parts IV.B.1-.2.

386. *Supra* Part IV.B.2.b; *see also supra* Part III.B.2.f.

387. *Supra* Part IV.B.1.a.

388. *Supra* Part III.B.2.c.

389. *Supra* Part III.B.2.e.

390. *See supra* Part III.B.2.c.

391. Hassija et al., *supra* note 294, at 48 ("The degree to which a person can comprehend and foresee the results of an ML model is known as interpretability."); Guidotti et al., *supra* note 294, at 6 (discussing interpretability in terms of "extent the model and/or its predictions are human understandable").

392. *See supra* Part III.B.2.e.

If, like the PlumsUp programmer, she wrote the rules underlying the PlumGPT model, there would be no doubt that she knows how it works and could anticipate the output with certainty. But if she didn't write the rules, the only way she could know with certainty what the output would be is if she could comprehensively explain how it works. Thus, while the PlumGPT programmer could theoretically achieve speech certainty through a comprehensive understanding of her algorithm's operations, the nature of machine learning models makes this impossible.[393] The relationships between all the features analyzed within the PlumGPT model are too complex for her to understand and too voluminous for her to deconstruct after the fact.[394]

Because the PlumGPT programmer didn't write the rules, nor can she explain how the PlumGPT model works, she cannot know with certainty what its output—her "speech"—will be when it is generated.

The necessary conclusion is that because machine learning models are probabilistic, they can never achieve 100% certainty in their predictions;[395] and because the programmer neither wrote the rules nor can comprehensively explain or understand them, the programmer can never know with certainty what the output of her model will be. To some degree, large or small, there will always be false negatives (a plum identified as a peach) and false positives (a peach identified as a plum).[396] And the programmer can never be sure when or why they will occur. She cannot know what her algorithmic speech will be at the moment the algorithm generates that speech. The output of machine learning algorithms therefore lacks speech certainty and falls outside the scope of the First Amendment's protection.

b. Machine learning output is not protected editorial discretion because it lacks speech certainty

The output of a machine learning model is protected under the doctrine of editorial discretion only if the model's use to determine a post's fitness for

393. MOLNAR, *supra* note 297, at ch. 3.3.2. ("Any model that exceeds a handful of parameters or weights is unlikely to fit into the short-term memory of the average human."); Jane C. Ginsburg & Luke Ali Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L.J. 343, 402 (2019) ("[Programming] quickly become[s] too complex and multi-dimensional for human programmers to comprehend.").

394. *See supra* Part III.B.2.e.

395. Loukides, *supra* note 4 ("[W]e have to be aware that machine learning is never going to be a 100% solution . . . ."); *see also* MICHAEL J. EVANS & JEFFREY S. ROSENTHAL, PROBABILITY AND STATISTICS: THE SCIENCE OF UNCERTAINTY 1 (2d ed. 2023) ("Probability is the science of uncertainty.").

396. *See* Loukides, *supra* note 4.

publication qualifies as an exercise of the programmer's judgment as to the contents of the published compilation.[397] As we explain below, it does not.

The analysis of the PlumGPT algorithm differs starkly from that in the first two hypotheticals. To understand why, try to find where the programmer makes the protected "selection and choice of material" that will or will not be included on her platform.[398] Where can we find the "human editorial judgments . . . responsible for producing the speech"?[399] With traditional programming, those judgments will be spelled out right in the code.[400] In PlumGPT's case, however, the programmer articulates no reasoning at all in the traditional code apart from deference to the machine learning model.[401] On the surface, the PlumGPT programmer appears to be deferring to a statistical analysis in the same way as the PlumsUp programmer deferred to a statistical analysis of an image's colors.[402] But the PlumsUp programmer deferred to traditional code; that is, her own judgments translated directly into programming language.[403]

The PlumGPT programmer, however, defers to gradient descent; that is, the mathematical process behind machine learning that empowers it to generate predictions *without* the programmers' judgment as to how those predictions should be made.[404] As explained above, that task was executed by gradient descent when it determined the model's parameters.[405] Because these parameters directly determine whether an image is fit for publication on PlumGPT, they constitute the relevant judgment for the First Amendment analysis. And, as explained in depth in Part III and more briefly above, that judgment cannot in any meaningful way be said to be the programmer's.[406]

Thus, the editorial discretion analysis directly follows from the speech certainty analysis. Because the machine learning programmer did not write the

397. *See supra* Part II.A.

398. Columbia Broad. Sys., Inc. v. Democratic Nat'l. Comm., 412 U.S. 94, 124 (1973).

399. VOLOKH & FALK, *supra* note 315, at 11.

400. Let's return to our traditional code hypo: "If a user attempts to publish a post containing the word 'eggplant,' serve the user with an error message reading, 'Sorry, this is an eggplant-free zone.' Otherwise, publish the post to the MeatUp feed." *See supra* Part IV.B.1.

401. *Supra* Part IV.B.3 ("[T]he PlumGPT model can identify whether a photo contains a plum with an 88% success rate . . . If the PlumGPT model determines the image is a plum, publish it.").

402. *Supra* Part IV.B.2 ("If a user attempts to publish a photo, analyze the colors of the image, and if at least 70% of the image is purple, publish it.").

403. *See supra* Part III.B.1.

404. *See supra* Part III.B.2.c.

405. *See supra* Part III.B.2.c.

406. *See supra* Part III.B.2.c.

model's rules and cannot comprehensively explain or understand them, the programmer can therefore never be certain what the output of the algorithm will be at the moment it is generated.[407] She cannot know with certainty that the PlumGPT feed will include what she intended for it to include and exclude the rest.[408] In other words, because the output of a machine learning model lacks speech certainty, it cannot meet the requirements of protected editorial discretion.

c. Machine-learning output also doesn't qualify as expressive conduct because it lacks speech certainty

Finally, the output of the PlumGPT model fails both prongs of the *Spence* test and cannot be appropriately characterized as the programmer's expressive conduct.

First, the PlumGPT programmer cannot have the requisite intent to convey a particularized message through her conduct when she cannot be certain what her conduct will be.[409] For the PlumGPT model, the purportedly expressive conduct is the act of publishing a photo of a plum or refusing to publish a photo of a non-plum. As described above, because of the probabilistic nature of machine learning algorithms, the PlumGPT algorithm will publish non-plums 12% of the time. But unlike the probabilistic traditional code of the PlumsUp programmer, the PlumGPT programmer cannot know when or why her machine learning algorithm will do so.[410] In other words, she can never be certain as to what her purported conduct will be.

As to the PlumGPT programmer's intended message, it may be that she wants to communicate the platform's disapproval of non-plum fruits through the algorithm.[411] But for purposes of the *Spence* test, that intent is only evinced in tandem with the conduct[412]—namely, how the machine learning model will sort a given photo. If the programmer has *any* doubt as to the output, then the alignment between any intended message and the accompanying conduct is essentially reduced to coincidence. The PlumsUp programmer, who writes the rules underlying her traditional code, has no such doubt.[413] But the PlumGPT programmer, who neither writes the rules, nor can explain them, cannot

407. *Supra* Part III.B.2.e.

408. *Supra* Part III.B.2.f.

409. *Supra* Part II.B.1.

410. *Supra* Part IV.B.3.b.

411. *See supra* Part IV.B.3.

412. *See supra* Part IV.B.3.

413. *Supra* Parts IV.B.1-2.

escape this doubt.[414] As a result, her intended message will align with her conduct some, but not all, of the time. And when it doesn't, it hardly makes sense to say that a speaker intended to communicate a message disapproving of non-plums by filtering out photos of non-plums when, in fact, she published a photo of a peach for reasons she can't explain.

In other words, the probabilistic nature of machine learning, in effect, requires programmers to rely on statistics for their intended message to align with their conduct. Extending protection to mistaken or accidental publication is a far cry from the "intentional" message required under the *Spence* test.[415] Protection for expressive conduct cannot reasonably depend on chance.

The second prong of the *Spence* test fails for more obvious reasons. If the programmer doesn't have certainty in the output of the model—in other words, she cannot guarantee that the machine learning model will sort a given photo correctly—then surely a reasonable observer couldn't divine the message from the mistaken publication either.[416] A PlumGPT user whose feed includes photos of a plum and an eggplant would not, for example, reasonably understand the PlumGPT programmer's message to be anti-eggplant.[417]

Thus, the output of machine learning models fails both prongs of the *Spence* test because it is not characterized by speech certainty. Precisely because the programmer cannot know with certainty what the output will be at the moment the algorithm generates it, she cannot have the requisite intent to communicate her intended message in tandem with that conduct; nor can a

414. The key difference between these two methods of programming is speech certainty. In probabilistic traditional code, the programmer knows exactly what the output of the model will be—any photo with 70% purple hues will be sorted as a plum. But in the machine-learning predictive model, the programmer cannot know how any given photo will be sorted and what parameters will be most decisive. Thus, unlike the traditional programmer, the mistakes aren't "intentional" insofar as the programmer can know why they would result.

415. *See supra* Part II.B.1.

416. *See supra* Part II.B.2.

417. We have assumed that each publication of a compilation of posts to a user is an individual act of expressive conduct. Another way one might conceive of algorithmic output is as a continuous and ongoing act of expressive conduct across all publications of all compilations of posts to all users. In that conception, instead of analyzing the machine-learning output as expressive conduct on a publication-by-publication basis (e.g., each published feed, or even published post, is an act of expressive conduct), one might insist that the expressive conduct is the continuous, aggregate output of the algorithm. Thus, one wouldn't judge whether a reasonable observer could ascertain the programmer's intended message from a single instance of publication, but rather the whole panoply of publications across all of the platform's user timelines. Admittedly, this may be a successful workaround for clearing the second prong of the *Spence* test. But even under this conception, the purportedly expressive conduct couldn't overcome the lack of speech certainty inherent in the algorithm's output in the first instance. *See supra* Part IV.A.

reasonable observer understand what that intended message is when it inevitably and inexplicably presents the observer with content that contradicts the intended message.[418]

* * *

In both the MeatUp and PlumsUp hypotheticals, the programmers' use of traditional code guarantees that they will never "publish that which reason tells them should not be published."[419] Although they used algorithms of different complexity, both programmers articulated their reasoning in the code and can be certain that this reasoning will be faithfully reflected in their algorithms' output. This guarantee both ensures their published compilations are characterized by speech certainty and protects those compilations as speech under the doctrine of editorial discretion. The speech certainty inherent in traditional code also likely qualifies the MeatUp and PlumsUp platforms as expressive conduct.[420] The PlumGPT hypothetical, however, is where algorithmic speech ceases to be characterized by speech certainty. Because the machine learning algorithm determines its own rules to decide which posts are fit for publication, and the programmer cannot comprehensively explain or understand those rules, the programmer cannot know with certainty what the output will be when it is generated. The output of machine learning algorithms lacks speech certainty and therefore is not speech within the meaning of the First Amendment.

418. Importantly, any accompanying speech on the platform's page wouldn't change the outcome of the expressive conduct analysis for the output of a machine learning model. Even if, for example, the community guidelines strictly prohibit the posting of any non-plum photos, there is a long description on the website's "About Us" page extolling the superiority of plums in the fruit kingdom, and the name of the website is OnlyPlums.com, explanatory speech cannot convert the machine learning algorithm's output into expressive conduct covered by the First Amendment. Rumsfeld v. F. for Acad. & Institutional Rts., Inc., 547 U.S. 47, 66 (2006) (explaining that explanatory speech doesn't render an action expressive). The proper unit of analysis in expressive conduct cases is the action that the individual claims as her speech, not the speech that accompanies it. Here then, the programmer's speech is for purposes of the expressive conduct inquiry is the output of the algorithm, not the explanatory words that appear elsewhere on her website. The Supreme Court has indicated that accompanying speech is not simply neutral in the analysis, but a death knell for the inquiry: The necessity of accompanying speech only underscores the non-expressive quality of the action. *Id.*

419. Mia. Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 256 (1974) (internal quotations omitted).

420. Although we note again that PlumsUp presents a closer question for the second prong of the *Spence* test. *Supra* Part IV.B.2.c.

V. Regulatory Implications

Recognition of the principle of speech certainty would place the output of machine learning algorithms beyond the reach of the First Amendment. State and federal governments would thus be free to regulate machine learning algorithms, including vast swaths of activity by social media platforms, search engines, and artificial intelligence companies. Against the common wisdom that all algorithmic output is protected speech, this is a jarring consequence. But we believe, on consideration, the shock ought to fade for most readers for three reasons.

First and foremost, whatever its consequences, the speech certainty principle is a historically and doctrinally rooted principle.[421] And while its consequences may be startling, the principle itself is surprisingly ordinary: Speech is that which a speaker knows with certainty he says when he says it. This definition of speech does not encompass the output of machine learning algorithms.[422] Rather than contorting the definition of speech to accommodate speech that a speaker doesn't know he says, we should be comfortable with the prospect of, after more than two centuries, at the dawn of a technological paradigm shift, having discovered the definition's limits. Moreover, the principle of speech certainty explains the emerging and widespread sense that machine learning algorithms are somehow different from what has come before it.[423] Speech certainty gives this intuition a principled foundation, rooted in First Amendment jurisprudence. Machine learning algorithms feel different because they *are* different—they run afoul of the speech certainty principle.

Second, the speech certainty principle does not alter any existing legal frameworks in any way. It leaves the First Amendment jurisprudence that has evolved over decades and centuries entirely undisturbed.[424] As we explained in Parts I and II, speech certainty has been an unspoken assumption of the First Amendment since the Founding and through to the present; no doctrine needs any reconsideration or modification to accommodate it. Nor does the speech

421. *See supra* Part I (explaining why the text, history, and purposes of the First Amendment compel recognition of the speech certainty principle); *supra* Part II (explaining why relevant First Amendment precedent does the same).

422. *See supra* Parts I-II (illustrating speech certainty's compatibility with First Amendment text, history, purposes, and relevant doctrine); *supra* Part IV (explaining why machine learning output falls outside the First Amendment's protection).

423. *See supra* note 16 (identifying emerging regulatory efforts to restrain the influence of social media platforms); Lessig, *supra* note 23, at 278 (noting that "technology has changed fundamentally").

424. *See supra* Parts I-II (illustrating speech certainty's compatibility with First Amendment text, history, purposes, and relevant doctrine); *see also supra* note 140 (addressing how the principle of speech certainty comports with *Smith v. California*, 361 U.S. 147 (1959)).

certainty principle alter adjacent legal frameworks, such as those that insulate platforms from liability and incentivize technological innovation. Nothing about the First Amendment status of machine learning algorithms affects the application of Section 230 of the Communications Decency Act to the social media platforms that employ them, for example.[425] Nor does it touch on the whole panoply of intellectual property doctrines and statutes that, for example, protect platforms' algorithms and their outputs from commercial misappropriation.[426] Recognition of the speech certainty principle—which only defines "speech" under the First Amendment—would have no legal effect apart from removing machine learning output from the First Amendment's ambit.

Finally, while the speech certainty principle does place machine learning algorithms outside the protection of the First Amendment, it does no more than that. It does not prescribe that machine learning algorithms be regulated by the government in any particular way, or even that they be regulated at all. Instead, it grants our democracy the freedom to decide what, if anything, to do with these powerful new technologies that are poised to reshape our world in the coming years. Rather than allow platforms to forestall regulation of their machine learning algorithms by invoking their First Amendment rights on questionable legal ground, the speech certainty principle frees up space for debate, leaving the biggest questions about regulation of algorithms to the democratic process.

Conclusion

The speech certainty principle is the simple idea that if you don't know what you're saying when you say it, then whatever you said isn't your "speech" within the meaning of the First Amendment. At the Founding, when only oral, written, and printed speech was possible, all speech necessarily fit within that understanding.[427] Since then, communications technology evolved, giving way to the telegraph, radio, television, and the internet. But although speech could now be transmitted across vast distances, instantaneously and en masse, the

425. 47 U.S.C. § 230.

426. Platforms have successfully argued that the misappropriation of algorithmic source code or training models violates trade secret law. *See, e.g.,* LivePerson, Inc. v. 24/7 Customer, Inc., 83 F. Supp. 3d 501, 514-15 (S.D.N.Y. 2015) (finding algorithms based on artificial intelligence eligible for trade secret protection). In some circumstances, the source code of machine learning algorithms might also be granted copyright protection to achieve the same ends. The speech certainty principle—which only defines "speech" under the First Amendment—would have little bearing on the application of such doctrines to machine learning algorithms.

427. *Supra* Parts I.A-.B.

speech certainty principle held.[428] In any medium, the speaker always knew what she said when she said it. Over centuries, this elemental feature of speech therefore revealed itself as a cornerstone of First Amendment jurisprudence, most recently in the doctrines of editorial discretion and expressive conduct.[429]

It may be surprising that such a foundational requirement underlying such a foundational freedom has gone unspoken until now. But for all history—until roughly the last decade—speech simply could not avoid being imbued with speech certainty. Only as the internet gave way to online platforms, which in turn gave way to machine learning algorithms, has the prospect of "speech" without speech certainty become possible.[430] Unlike previous algorithms written with traditional code, machine learning algorithms write their own rules, which their human programmers can neither fully explain nor comprehend.[431] The rapid rise of these algorithms thus raises vital questions about if and how to draw the line when human speech morphs into machine "speech."[432]

To date, scholars have generally balked at the possibility of drawing such lines.[433] We believe, however, that the speech certainty principle provides a coherent, principled response. Speech certainty isn't a departure from current First Amendment jurisprudence. It's the logical continuation of First Amendment doctrine that has only ever protected speech which the speaker can be certain she said. Indeed, failing to at least consider the speech certainty principle could result in the inadvertent extension of First Amendment protection to "speech" which the speaker doesn't know he's "said."

In *Moody v. Netchoice*, the Supreme Court tentatively ventured that "*some* platforms, in at least *some* functions, are indeed engaged in expression."[434] But it expressly and repeatedly put an asterisk on that conclusion: it was based only on the existing, undeveloped record. Further development of that record will

428. *Supra* Part II.A (explaining how editorial discretion comports with the principle of speech certainty across different media).

429. *Supra* Parts I-II (illustrating speech certainty's compatibility with First Amendment text, history, purposes, and relevant doctrine).

430. *See supra* Part III (describing the transition from traditional code to machine learning); *see also supra* Parts IV.B.1.a, IV.B.2.a, and IV.B.3.a (analyzing whether traditional code and machine learning are "speech" under the First Amendment).

431. *Supra* Part III.B.2.e.

432. Lessig, *supra* note 23, at 276 ("At some point along the continuum between your first program, 'Hello world!' and [artificial intelligence], the speech of machines crosses over from speech properly attributable to the coders to speech no longer attributable to the coders.").

433. *See, e.g.*, Benjamin, *supra* note 112, at 1448 (fearing it "will entail a radical revamping of our Free Speech Clause jurisprudence"); Volokh et al., *supra* note 338, at 653 (cautioning against "murky line-drawing").

434. 144 S. Ct. 2383, 2393 (2024) (emphasis added).

show in fact what this Article has explored in theory: that the machine learning models on which social-media platforms rely to rank, recommend, and remove content on their feeds does not match the definition of editorial discretion as the Court articulated in *Moody*.[435] Instead, it will show that the platforms can never be certain that the content published by those models will align with what they intended to publish.[436] In fact, because these probabilistic machine learning models will always be wrong at least some of the time, it is guaranteed that the platforms will publish precisely what they intended not to publish.[437] In other words, the platforms' algorithmic output lacks speech certainty, and thus doesn't qualify as "speech" within the meaning of the First Amendment.[438]

By bringing speech certainty to the table as the Supreme Court is poised to shape the online speech environment in the coming years, our aim is to inform what could be a watershed moment for First Amendment doctrinal development. The growing prevalence of machine learning algorithms will, sooner or later, demand that we address the question about the First Amendment protection of their output. The speech certainty principle offers a historically and doctrinally rooted answer.

435. *Supra* note 6 (collecting sources).

436. *Compare supra* Part III.B.2 (explaining the mechanics of machine learning), *with* Part III.C (showing platforms' reliance on machine learning).

437. *Supra* Part III.B.2.e.

438. *Supra* Part IV.B.3.