WILL ARTIFICIAL INTELLIGENCE EAT THE LAW? THE RISE OF HYBRID SOCIAL-ORDERING SYSTEMS

*Tim Wu**

Software has partially or fully displaced many former human activities, such as catching speeders or flying airplanes, and proven itself able to surpass humans in certain contests, like Chess and Go. What are the prospects for the displacement of human courts as the centerpiece of legal decisionmaking? Based on the case study of hate speech control on major tech platforms, particularly on Twitter and Facebook, this Essay suggests displacement of human courts remains a distant prospect, but suggests that hybrid machine–human systems are the predictable future of legal adjudication, and that there lies some hope in that combination, if done well.

INTRODUCTION

Many of the developments that go under the banner of artificial intelligence that matter to the legal system are not so much new means of breaking the law but of bypassing it as a means of enforcing rules and resolving disputes.[1] Hence a major challenge that courts and the legal system will face over the next few decades is not only the normal challenge posed by hard cases but also the more existential challenge of supersession.[2]

Here are a few examples. The control of forbidden speech in major fora, if once the domain of law and courts, has been moving to algorithmic judgment in the first instance.[3] Speeding is widely detected and punished by software.[4] Much enforcement of the intellectual property

* Julius Silver Professor of Law & Technology, Columbia Law School. My thanks to Vince Blasi, Ryan Doerfler, Noah Feldman, Sarah Knuckey, David Pozen, Olivier Sylvain, and participants in the Columbia faculty workshop series.

1. In this Essay, the broader meaning of "artificial intelligence" is used—namely a computer system that is "able to perform tasks normally requiring human intelligence" such as decisionmaking. Artificial Intelligence, Lexico, https://www.lexico.com/en/definition/artificial_intelligence [https://perma.cc/86XR-2JZ8] (last visited July 31, 2019).

2. A small legal literature on these problems is emerging. See, e.g., Michael A. Livermore, Rule by Rules, *in* Computational Legal Studies: The Promise and Challenge of Data-Driven Legal Research (Ryan Whalen, ed.) (forthcoming 2019) (manuscript at 1–2), https://papers.ssrn.com/abstract=3387701 (on file with the *Columbia Law Review*); Richard M. Re & Alicia Solow-Niederman, Developing Artificially Intelligent Justice, 22 Stan. Tech. L. Rev. 242, 247–62 (2019); Eugene Volokh, Chief Justice Robots, 68 Duke L.J. 1135, 1147–48 (2019).

3. See infra Part II.

4. See Jeffrey A. Parness, Beyond Red Light Enforcement Against the Guilty but Innocent: Local Regulations of Secondary Culprits, 47 Willamette L. Rev. 259, 259 (2011) ("Automated traffic enforcement schemes, employing speed, and red light cameras, are increasingly used by local governments in the United States." (footnote omitted)).

laws is already automated through encryption, copy protection, and automated takedowns.[5] Public prices once set by agencies (like taxi prices) are often effectively supplanted by prices set by algorithm.[6] Blockchain agreements are beginning to offer an alternative mechanism to contract law for the forging of enforceable agreements.[7] Software already plays a role in bail determination and sentencing,[8] and some are asking whether software will replace lawyers for writing briefs, and perhaps even replace judges.[9]

Are human courts just hanging on for a few decades until the software gets better? Some might think so, yet at many points in Anglo American legal history, courts have been thought obsolete, only to maintain their central role. There are, it turns out, advantages to adjudication as a form of social ordering that are difficult to replicate by any known means.[10] This Essay predicts, even in areas in which software has begun to govern, that human courts[11] will persist or be necessarily reinvented. It predicts, however, that human–machine hybrids will be the first replacement for human-only legal systems, and suggests, if done right, that there lies real promise in that approach. The case study of content control on online platforms and Facebook's review board is used to support these descriptive and normative claims.

The prediction that courts won't wholly disappear may seem an easy one, but what's more interesting is to ask why, when software is "eating" so many other areas of human endeavor. Compared with the legal system, software has enormous advantages of scale and efficacy of enforcement. It might tirelessly handle billions if not trillions of decisions in the time it takes a human court to decide a single case. And even more importantly, the design of software can function as an ex ante means of ordering that does not suffer the imperfections of law enforcement.[12]

But human courts have their own advantages. One set of advantages, more obvious if perhaps more fragile, is related to procedural fairness. As between a decision made via software and court adjudication, the latter, even if delivering the same results, may yield deeper acceptance and

5. See infra Part I.

6. See Jessica Leber, The Secrets of Uber's Mysterious Surge Pricing Algorithm, Revealed, Fast Company (Oct. 29, 2015), https://www.fastcompany.com/3052703/the-secrets-of-ubers-mysterious-surge-pricing-algorithm-revealed [https://perma.cc/H7MB-SC8T].

7. See Eric Talley & Tim Wu, What Is Blockchain Good for? 5–9 (Feb. 28, 2018) (unpublished manuscript) (on file with the *Columbia Law Review*).

8. See Algorithms in the Criminal Justice System, Elec. Privacy Info. Ctr., https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/KY2D-NQ2J] (last visited Apr. 23, 2019).

9. See Volokh, supra note 2, at 1144–48, 1156–61.

10. See Lon L. Fuller, The Forms and Limits of Adjudication, 92 Harv. L. Rev. 353, 357 (1978).

11. This Essay uses "courts" to refer to any adjudicative body, public or private, that resolves a dispute after hearing reasoned argument and explains the basis of its decision.

12. Lawrence Lessig, Code: Version 2.0, at 124–25 (2006).

greater public satisfaction.[13] In the future, the very fact of human decision—especially when the stakes are high—may become a mark of fairness.[14] That said, society has gotten used to software's replacement of humans in other areas, such as the booking of travel or the buying and selling of stocks, so this advantage may be fragile.

A second, arguably more lasting advantage lies in human adjudication itself and its facility for "hard cases" that arise even in rule-based systems. Most systems of social ordering consist of rules, and a decisional system that was merely about obeying rules might be replaced by software quite easily. But real systems of human ordering, even those based on rules, aren't like that.[15] Instead, disputes tend to be comprised of both "easy cases"—those covered by settled rules—and the aforementioned "hard cases"—disputes in which the boundaries of the rules become unclear, or where the rules contradict each other, or where enforcement of the rules implicates other principles.[16] There is often a subtle difference between the written rules and "real rules," as Karl N. Llewellyn put it.[17] Hence, a software system that is instructed to "follow the rules" will produce dangerous or absurd results.

Justice Cardozo argued that the judicial process "in its highest reaches is not discovery, but creation . . . [by] forces . . . seldom fully in consciousness."[18] Better results in hard cases may for a long time still depend instead on accessing something that remains, for now, human—that something variously known as moral reasoning, a sensitivity to evolving norms, or a pragmatic assessment of what works. It is, in any case, best expressed by the idea of exercising "judgment." And if the courts do indeed have a special sauce, that is it.

It is possible that even this special sauce will, in time, be replicated by software, yielding different questions.[19] But as it stands, artificial intelligence (AI) systems have mainly succeeded in replicating human decisionmaking that involves following rules or pattern matching—chess

---

13. See generally E. Allan Lind & Tom R. Tyler, The Social Psychology of Procedural Justice (Melvin J. Lerner ed., 1988) (providing a classic overview of the advantages human courts have over software because of procedural fairness considerations in human courts, potentially ignored by software, that lead to acceptance of their decisions by the public).

14. See Aziz Z. Huq, A Right to a Human Decision, 105 Va. L. Rev. (forthcoming 2020) (manuscript at 21–22), https://ssrn.com/abstract=3382521 (on file with the *Columbia Law Review*) (explaining that concerns about transparency have led some to demand human decisions).

15. This refers to a caricatured version of H.L.A. Hart's view of what law is. See H.L.A. Hart, The Concept of Law 2–6 (Peter Cane, Tony Honoré & Jane Stapleton eds., 2d ed. 1961).

16. Cf. Ronald Dworkin, Hard Cases, *in* Taking Rights Seriously 81, 81 (1978) (stating that when "hard cases" fall on the edge of clear rules, judges have the discretion to decide the case either way).

17. Karl N. Llewellyn, The Theory of Rules 72–74 (Frederick Schauer ed., 2011).

18. Benjamin N. Cardozo, The Nature of the Judicial Process 166–67 (1921).

19. See Huq, supra note 14, at 18; Volokh, supra note 2, at 1166–67, 1183–84.

and Jeopardy! are two examples.[20] It would risk embarrassment to argue that machines will *never* be able to make or explain reasoned decisions in a legal context, but the challenges faced are not trivial or easily overcome.[21] And even if software gets better at understanding the nuances of language, it may still face the deeper, jurisprudential challenges described here. That suggests that, for the foreseeable future, software systems that aim to replace systems of social ordering will succeed best as human–machine hybrids, mixing scale and efficacy with human adjudication for hard cases. They will be, in an older argot, "cyborg" systems of social ordering.[22]

When we look around, it turns out that such hybrid systems are already common. Machines make the routine decisions while leaving the hard cases for humans. A good example is the flying of an airplane, which, measured by time at the controls, is now mostly done by computers, but sensitive, difficult, and emergency situations are left to a human pilot.[23]

The descriptive thesis of this Essay is supported by a case study of content control (the control of hate speech, obscenity, and other speech) on online platforms like Facebook and Twitter. Despite increasing automation, the generation of hard questions has yielded the development, by the major platforms, of deliberative bodies and systems of appeal, such as Facebook's prototype content review board, designed to rule on the hardest of speech-related questions. In the control of online speech, and in the autopilot, we may be glimpsing, for better or worse, the future of social ordering in advanced societies.

While automated justice may not sound appealing on its face, there is some real promise in the machine–human hybrid systems of social ordering described here. At their best, they would combine the scale and

---

20. See John Markoff, Computer Wins on 'Jeopardy!': Trivial, It's Not, N.Y. Times (Feb. 16, 2011), https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html (on file with the *Columbia Law Review*).

21. See Drew McDermott, Why Ethics Is a High Hurdle for AI 2 (2008), http://www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf [https://perma.cc/PZ7Z-5QFN]; Adam Elkus, How to Be Good: Why You Can't Teach Human Values to Artificial Intelligence, Slate (Apr. 20, 2016), https://slate.com/technology/2016/04/why-you-cant-teach-human-values-to-artificial-intelligence.html [https://perma.cc/4LR9-Q2GH]. But see IBM's project debater, a program that presents arguments on one side of an issue, and thereby might be thought to replicate aspects of lawyering. Project Debater, IBM, https://www.research.ibm.com/artificial-intelligence/project-debater/ [https://perma.cc/4A2W-XSB9] (last visited July 31, 2019).

22. The Merriam-Webster dictionary defines "cyborg" as "a bionic human," meaning a being comprised of mixed human and machine elements, like the fictional character "Darth Vader" from the twentieth-century film series *Star Wars*. See Cyborg, Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/cyborg [https://perma.cc/FD6A-UWZ2] (last visited Aug. 31, 2019).

23. Reem Nasr, Autopilot: What the System Can and Can't Do, CNBC (Mar. 27, 2015), https://www.cnbc.com/2015/03/26/autopilot-what-the-system-can-and-cant-do.html [https://perma.cc/Q7CT-GBLN].

effectiveness of software with the capacity of human courts to detect errors and humanize the operation of the legal system. As Lon Fuller put it, human courts are "a device which gives formal and institutional expression to the influence of reasoned argument in human affairs."[24] What might the court system look like without scaling problems—if routine decisions went to machines, reducing the court's own workload as a major factor in decisionmaking? To be sure, what could arise are inhumane, excessively rule-driven systems that include humans as mere tokens of legitimization,[25] but hopefully we can do better than that.

This Essay provides advice both for designers of important AI-decision systems and for government courts. As for the former, many AI systems outside of the law have aspired to a complete replacement of the underlying humans (the self-driving car,[26] the chess-playing AI[27]). But when it comes to systems that replace the law, designers should be thinking harder about how best to combine the strengths of humans and machines, by understanding the human advantages of providing a sense of procedural fairness, explainability, and the deciding of hard cases. That suggests the deliberate preservation of mechanisms for resort to human adjudication (either public or private) as part of a long-term, sustainable system.

Human courts, meanwhile, should embark on a greater effort to automate the handling of routine cases and routine procedural matters, like the filing of motions. The use of intelligent software for matters like sentencing and bail—decisions with enormous impact on people's lives—seems exactly backward. The automation of routine procedure might help produce both a much faster legal system and also free up the scarce resource of highly trained human judgment to adjudicate the hard cases, or to determine which are the hard cases. Anyone who has worked in the courts knows that the judiciary's mental resources are squandered on thousands of routine matters; there is promise in a system that leaves judges to do what they do best: exercising judgment in the individual case, and humanizing and improving the written rules. This also implies that judges should seek to cultivate their comparative advantage, the exercise of human judgment, instead of trying to mimic machines that follow rules.[28]

---

24. Fuller, supra note 10, at 366.

25. This is a concern expressed in Re & Solow-Niederman, supra note 2, at 246–47.

26. Alex Davies, The WIRED Guide to Self-Driving Cars, WIRED (Dec. 13, 2018), https://www.wired.com/story/guide-self-driving-cars/ [https://perma.cc/7P3A-5NLE].

27. Natasha Regan & Matthew Sadler, DeepMinds's Superhuman AI Is Rewriting How We Play Chess, WIRED (Feb. 3, 2019), https://www.wired.co.uk/article/deepmind-ai-chess [https://perma.cc/WF5Z-VEV7].

28. Cf. Kathryn Judge, Judges and Judgment: In Praise of Instigators, 86 U. Chi. L. Rev. (forthcoming 2019) (manuscript at 1–2), https://ssrn.com/abstract=3333218 (on file with the *Columbia Law Review*) (describing Judge Richard Posner's rejection of such machine-like jurisprudence).

Part I frames software decisionmaking among its competitors as a system of social ordering. Part II introduces the case study of Twitter and Facebook's handling of hate speech, focusing on the evolution of online norms, and the subsequent adoption of a hybrid human–software system to control speech. Part III assesses, from a normative perspective, the hybrid systems described in Part II.

I. BACKGROUND

In his book *Empire and Communications*, Harold Innes sought to characterize civilizations by their primary medium of communication.[29] The oral tradition of the ancient Greeks, he argued, informed the development of Greek philosophy; the Egyptian civilization changed as it transitioned from stone engraving to papyrus; medieval thought was shaped by the codex, and so on.[30] For our purposes, we need a different taxonomy of civilizations, one that characterizes societies not by medium but by how they make important public decisions (or, in Fuller's phrase, accomplish "social ordering").[31] By this I mean decisions that are both important and of public consequence, that define citizens' relationships with each other.

Under this way of seeing things, civilizations and societies really do differ. One axis is the divide between private and public bodies. Another is how much is governed by social norms as opposed to positive law.[32] Ordering might be more or less centralized; and there is the method of decision itself, which, as Fuller suggested, might be adjudicative, legislative, or accomplished by contractual negotiation.[33] I will not bother to pretend that the lines I have mentioned are the only ways you might imagine the division.[34]

This broader view demonstrates that decisional systems are in an implicit competition. Matters may drift between private and public decisionmaking, or between norms and law, and decisions can become more centralized or decentralized. Over the last 200 years, in the United States and commonwealth countries, a decentralized common law has been

---

29. See generally Harold Innes, Empire and Communications (David Godfrey ed., 1950) (arguing that communication provides crucial insight into a civilization's organization and administration of its government, and comparing various civilizations including Egypt and the Roman Empire based on their communication).

30. See id. at 5.

31. See Fuller, supra note 10, at 357.

32. See Robert C. Ellickson, Order Without Law: How Neighbors Settle Disputes 1–11 (1994) (discussing how order is often achieved without law).

33. See Fuller, supra note 10, at 363 (arguing that "adjudication should be viewed as a form of social ordering, as a way in which the relations of men to one another are governed and regulated").

34. There is, for example, also the question of how centralized or decentralized the systems of order are. Lawrence Lessig divided the universe of regulative forms into four: law, code, norms, and markets. Lessig, supra note 12, at 124–25.

somewhat (though not fully) displaced by a more centralized statutory law, and then further displaced by regulations and the administrative state.[35] Matters once thought purely private, like the firing of an employee or one's conduct in the workplace, have become subjects of public decision, while others once public, like the control of obscenity and other forbidden speech, are now mainly the province of private institutions. There is much complementarity in social ordering: a murderer may be shamed, fired, and imprisoned. But there is also competition, as for example, when new laws "crowd out" longstanding norms.

That idea that systems of public ordering might compete (or complement each other) is not new,[36] but what *is* new is the arrival of software and artificial intelligence as a major modality of public decisionmaking. As first predicted by Lawrence Lessig, what might have been thought to be important public decisions have either been displaced or are beginning to be displaced by software, in whole or in part.[37] It is a subtle displacement, because it is both private and unofficial, and advancing slowly, but it is happening nonetheless.

That idea of being ruled by intelligent software may sound radical but, as suggested in the Introduction, it is not hard to find examples in which software accomplishes what might previously be described as public decisionmaking. A good example is the dissemination and reproduction of expressive works. The posting of copyrighted works on YouTube and other online video sites was once directly and actively governed by section 512 of the copyright code.[38] In a technical sense the law still governs, but over the last decade sites like YouTube have begun using software (named "Content ID") to intelligently and proactively take down copyrighted works.[39] This understanding, implemented in code, was undertaken in the shadow of the law, but it is not compelled by it, and the decisions made by the software are now more important than the law. In the criminal law, software has become an aid to decisionmaking, and sometimes the decisionmaker in some jurisdictions, for matters like

35. See Lawrence Friedman, A History of American Law 253–78, 503–15 (3d ed. 2005).

36. See, e.g., Emanuela Carbonara, Law and Social Norms, *in* 1 The Oxford Handbook of Law & Economics 466, 475–80 (noting that while legal norms can reinforce social norms by "bending them towards the law when discrepancy exists and favoring their creation where social norms do not exist," legal regulation can also "destroy existing social norms"); Yoshinobu Zasu, Sanctions by Social Norms and the Law: Substitutes or Complements?, 36 J. Legal Stud. 379, 379–82 (2007) (discussing whether informal sanctions imposed through social norms are in competition with, or complement, the formal sanctions of the law).

37. Lessig, supra note 12, at 125–37.

38. 17 U.S.C. § 512 (2012).

39. How Content ID Works, YouTube Help, https://support.google.com/youtube/answer/2797370?hl=en [https://perma.cc/CLD9-L2VK] [hereinafter YouTube Help, How Content ID Works] (last visited July 31, 2019).

setting bail or sentencing.[40] And as we shall see in much more detail below, the control of forbidden speech in major online fora is heavily dependent on software decisions.

To be sure, software remains in the early stages of replacing the law, and much remains completely outside software's ambit. But let us assume that software is at least beginning to be the method by which at least some decisions once made by the law are now made.[41] If that is true, then the best glimpse of what the future will hold lies in the systems that control offensive, hateful, and harmful speech online.

II. THE CASE STUDY: FACEBOOK, TWITTER, AND HEALTHY SPEECH ENVIRONMENTS

This Part provides a case study of the migration of Facebook and Twitter toward a norm of healthy speech environments and their implementation of such norms in hybrid systems of code and human judgment.

A. *The Evolution of Online Speech Norms from the 1990s Through 2016*

When the first online communities began emerging in the late 1980s and early 1990s, a widespread and early aspiration was the creation of public platforms that were "open and free" in matters of speech.[42] That desire reflected, in part, the "cyber-libertarian" tendencies among the pioneers of online technologies of that era.[43] The World Wide Web, which became popular in the early 1990s, was the chief enabling technology for the promise of a non-intermediated publishing platform for the masses. To a degree rarely, if ever, attempted in human history, the web and its major fora and platforms adhered to an "open and free" philosophy.[44] The Usenet, an early and public online discussion forum,

---

40. Algorithms in the Criminal Justice System, supra note 8 (listing different states' uses of algorithmic tools for sentencing, probation, and parole decisions).

41. The key to accepting this conclusion is to accede to the premise that the software is making decisions, which some may dispute. Some might ask if the software is really "deciding," as opposed to the programmer of the algorithm. I address these complications in Tim Wu, Machine Speech, 161 U. Pa. L. Rev. 1495 (2013).

42. Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads 252 (2016) [hereinafter Wu, Attention Merchants].

43. See Tim Wu & Jack Goldsmith, Who Controls the Internet: Illusions of a Borderless World 1–10 (2006).

44. Even in the 1990s, the online communities that experimented with purely laissez faire speech platforms ran into problems linked to trolling and abuse, and very few of the communities were completely without rules. See Wu, Attention Merchants, supra note 42, at 276–88. It is also true that, by the early 2000s, the law and courts began to demand compliance with their laws, including the copyright laws, drug laws, laws banning child pornography, and so on. See Wu & Goldsmith, supra note 43, at 65–85.

allowed any user to create their own forum on any topic.[45] The famous online chatrooms of the 1990s were largely uncensored.[46] MySpace, the most popular social networking platform before Facebook, allowed its users to use any name they wanted, and to say almost anything they wanted.[47]

The "open and free" ideal was aided by the enactment of Section 230 of the Communications Decency Act of 1996.[48] The law, which granted platform owners immunity from tort for content posted on their platforms, was described as a "good Samaritan" law to protect sites trying to take down offensive content.[49] In practice, and in judicial interpretation, section 230 granted blanket immunity to all online platforms, both good Samaritans and bad, thereby protecting those who followed an "anything goes" mentality.[50]

The "open and free" speech ideal remained an aspired-to norm for the first twenty years of the popular internet. But under pressure, it began to change decisively over the years 2016 and 2017.[51] It has been replaced with a widespread if not universal emphasis among the major platforms—especially Twitter and Facebook—on creating "healthy" and "safe" speech environments online.[52] To be sure, the change in norms

---

45. See Sandra L. Emerson, Usenet: A Bulletin Board for Unix Users, Byte, Oct. 1983, at 219, 219, https://archive.org/stream/byte-magazine-1983-10/1983_10_BYTE_08-10_UNIX#page/n219/mode/2up (on file with the *Columbia Law Review*).

46. See Wu, Attention Merchants, supra note 42, at 202–05; see also EJ Dickson, My First Time with Cybersex, Kernel (Oct. 5, 2014), https://kernelmag.dailydot.com/issue-sections/headline-story/10466/aol-instant-messenger-cybersex/ [https://perma.cc/98EC-6F3H] (recounting experiences with cybersex as a ten-year-old).

47. See Saul Hansell, For MySpace, Making Friends Was Easy. Big Profit Is Tougher., N.Y. Times (Apr. 23, 2006), https://www.nytimes.com/2006/04/23/business/yourmoney/23myspace.html (on file with the *Columbia Law Review*) (describing MySpace as "very open to frank discussion, provocative images and links to all sorts of activities" including profiles maintained by *Playboy* magazine and porn star Jenna Jameson); Michael Arrington, MySpace Quietly Begins Encouraging Users to Use Their Real Names, TechCrunch (Dec. 17, 2008), https://techcrunch.com/2008/12/17/myspace-quietly-begins-encouraging-users-to-use-their-real-names/ [https://perma.cc/CKR6-MLYZ] (noting pre-2009 anonymity of MySpace profiles).

48. Communications Decency Act of 1996, 47 U.S.C. § 230 (2012) (stating, among other things, that "[i]t is the policy of the United States . . . to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services").

49. See Andrew M. Sevanian, Section 230 of the Communications Decency Act: A "Good Samaritan" Law Without the Requirement of Acting as a "Good Samaritan," 21 UCLA Ent. L. Rev. 121, 125 (2014).

50. See Danielle Keats Citron & Benjamin Wittes, The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity, 86 Fordham L. Rev. 401, 413 (2017) ("An overbroad reading of the [Communications Decency Act] has given online platforms a free pass to ignore illegal activities, to deliberately repost illegal material, and to solicit unlawful activities while ensuring that abusers cannot be identified.").

51. See infra text accompanying notes 61–63.

52. See infra notes 61–63 and accompanying text.

has never been explicitly stated as such; but it is hard to deny the change in emphasis is not also a change in substance. We might put it this way: If the major American online platforms once (from the 1990s through the mid-2010s) tended to follow speech norms that generally resembled the First Amendment's, the new focus on healthy speech and acceptance of the concept of harmful speech is far closer to the European speech tradition and its bans on hate speech.[53]

What changed? The mid-2010s shift in online speech norms on major platforms can be understood as reflecting three major developments. The first has been the relative success of a broader social movement stressing the importance of "safe" environments, reflected most strongly at American college campuses in the 2010s.[54] Those norms began to spill over into increasingly strong critiques of the major internet speech platforms. By the mid-2010s, journalists and civil rights groups, for example, heavily criticized Twitter and Facebook for tolerating attacks on women and historically disadvantaged groups and thereby creating "toxic" spaces for its users.[55] As a Buzzfeed journalist wrote in 2016, "Twitter is as infamous today for being as toxic as it is famous for being revolutionary."[56]

A second reason was a political concern: a widespread perception that the platforms had tolerated so much dissemination of hateful speech, foreign interference with elections, atrocity propaganda, and hoaxes as to become a threat to democratic institutions. This critique

---

53. See Jeremy Waldron, The Harm in Hate Speech 6–17 (2012) (summarizing legal and philosophical differences between European and American speech traditions).

54. In 2015, a large survey found about 71% of college entrants agreed that "colleges should prohibit racist/sexist speech on campus." Kevin Eagan, Ellen Bara Stolzenberg, Abigail K. Bates, Melissa C. Aragon, Maria Ramirez Suchard & Cecilia Rios-Aguilar, Higher Educ. Research Inst., The American Freshman: National Norms Fall 2015, at 47 (2016), https://www.heri.ucla.edu/monographs/TheAmericanFreshman2015.pdf [https://perma.cc/82CJ-T53B].

55. See, e.g., Emily Dreyfuss, Twitter Is Indeed Toxic for Women, Amnesty Report Says, WIRED (Dec. 18, 2018), https://www.wired.com/story/amnesty-report-twitter-abuse-women/ [https://perma.cc/2NXS-SULS] (detailing high rates of abusive tweets directed toward women journalists and politicians); Robinson Meyer, The Existential Crisis of Public Life Online, Atlantic (Oct. 30, 2014), https://www.theatlantic.com/technology/archive/2014/10/the-existential-crisis-of-public-life-online/382017/ [https://perma.cc/7845-2PHH] (criticizing Twitter's lack of response to Gamergate); Hamza Shaban & Taylor Telford, Facebook and Twitter Get an Avalanche of Criticism About Russian Interference, L.A. Times (Dec. 18, 2018), https://www.latimes.com/business/technology/la-fi-tn-facebook-twitter-20181218-story.html [https://perma.cc/KS68-44S9] (describing the NAACP's criticism of Facebook for "the spread of misinformation and the utilization of Facebook for propaganda promoting disingenuous portrayals of the African American community").

56. Charlie Warzel, "A Honeypot for Assholes": Inside Twitter's 10-Year Failure to Stop Harassment, BuzzFeed News (Aug. 11, 2016), https://www.buzzfeednews.com/article/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s [https://perma.cc/V92P-NB7N].

emerged strongly after the 2016 election.[57] Relatedly, outside the United States over this period, Facebook faced heated accusations that its site was used to organize and promote violence in countries like Myanmar, Sri Lanka, and India.[58]

A final development was the ability, given consolidation in the speech platform market, for a limited number of platforms—Twitter, Facebook, Google—to have system-wide effects. (These platforms, all private actors, are of course unconstrained by constitutional norms.[59]) To be sure, there are some platforms, like 4chan, that remain devoted to the older laissez faire norm,[60] and specialized sites, like pornographic sites, that obviously take different views of sex and nudity. But by 2016, the major platforms, surely comprising most of the online speech in the world, had all effectively moved to treat hateful speech as potentially "violent," an "attack," and subject to removal.[61] The new norms of online speech are codified in the lengthy and highly specific content rules kept by Facebook, Google, and YouTube.[62] Simply put, they now regard many categories of speech as subject to removal, from the more easily defined (videos of suicide attempts, child pornography) to the more ambiguous

57. See, e.g., Alexis C. Madrigal, What Facebook Did to American Democracy, Atlantic (Oct. 12, 2017), https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/ [https://perma.cc/8AV5-AE3B] (chronicling Facebook's role in the 2016 elections and concluding that the "roots of the electoral system—the news people see, the events they think happened, the information they digest—had been destabilized").

58. See, e.g., Vindu Goel & Shaikh Azizur Rahman, When Rohingya Refugees Fled to India, Hate on Facebook Followed, N.Y. Times (June 14, 2019), https://www.nytimes.com/2019/06/14/technology/facebook-hate-speech-rohingya-india.html (on file with the *Columbia Law Review*); Paul Mozur, A Genocide Incited on Facebook, with Posts from Myanmar's Military, N.Y. Times (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html (on file with the *Columbia Law Review*); Amalini De Sayrah, Opinion, Facebook Helped Foment Anti-Muslim Violence in Sri Lanka. What Now?, Guardian (May 5, 2018), https://www.theguardian.com/commentisfree/2018/may/05/facebook-anti-muslim-violence-sri-lanka [https://perma.cc/Y4X2-YCAG].

59. See, e.g., Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1930 (2019) ("[W]hen a private entity provides a forum for speech, the private entity is not ordinarily constrained by the First Amendment because the private entity is not a state actor. The private entity may thus exercise editorial discretion over the speech and speakers in the forum.").

60. See generally Rules, 4chan, https://www.4chan.org/rules [https://perma.cc/MSQ7-PN7N] (last visited July 30, 2019) (designating spaces where racism, pornography, and grotesque violence are allowed).

61. See, e.g., Community Standards, Facebook, https://www.facebook.com/communitystandards/ [https://perma.cc/D27N-XJEY] (last visited July 30, 2019); Hate Speech Policy, YouTube Help, https://support.google.com/youtube/answer/2801939?hl=en [https://perma.cc/AZD2-VH4V] (last visited July 30, 2019).

62. See supra note 61.

(hate speech, dehumanizing speech, advocacy of violence or terrorism).[63]

The easiest way to see the change in norms is by observing the changes in language used by representatives of the major companies. In 2012, Twitter executives had described the firm as belonging to "the free speech wing of the free speech party" and suggested that, in general "we remain neutral as to the content."[64] Alexander Macgillivray, Twitter's general counsel at the time, regularly litigated subpoena requests, telling the press that "[w]e value the reputation we have for defending and respecting the user's voice . . . . We think it's important to our company and the way users think about whether to use Twitter, as compared to other services."[65]

In contrast, by the later 2010s, Twitter had begun to emphasize "health" and "safety" as primary concerns.[66] In an interview, Twitter CEO Jack Dorsey suggested the "free speech wing" quote "was never a mission of the company" and that "[i]t was a joke, because of how people found themselves in the spectrum."[67] And, as the official Twitter blog stated in 2017:

> Making Twitter a safer place is our primary focus. We stand for freedom of expression and people being able to see all sides of any topic. That's put in jeopardy when abuse and harassment stifle and silence those voices. We won't tolerate it and we're launching new efforts to stop it.[68]

There are many more examples. Microsoft President Brad Smith in 2018 opined that "we should work to foster a healthier online environment more broadly. . . . [D]igital discourse is sometimes increasingly toxic. There are too many days when online commentary brings out the worst in people."[69] And testifying before Congress in 2018, Facebook CEO

---

63. See, e.g., Objectionable Content, Community Standards, Facebook, https://www.facebook.com/communitystandards/objectionable_content [https://perma.cc/9TMH-R2HG] (last visited July 30, 2019).

64. Josh Halliday, Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party,' Guardian (Mar. 22, 2012), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/75Z2-NBZP].

65. Somini Sengupta, Twitter's Free Speech Defender, N.Y. Times (Sept. 2, 2012), https://www.nytimes.com/2012/09/03/technology/twitter-chief-lawyer-alexander-macgillivray-defender-free-speech.html (on file with the *Columbia Law Review*).

66. See Nicholas Thompson, Jack Dorsey on Twitter's Role in Free Speech and Filter Bubbles, WIRED (Oct. 16, 2018), https://www.wired.com/story/jack-dorsey-twitters-role-free-speech-filter-bubbles/ [https://perma.cc/D6HJ-HJTQ].

67. See id.

68. Ed Ho, An Update on Safety, Twitter: Blog (Feb. 7, 2017), https://blog.twitter.com/en_us/topics/product/2017/an-update-on-safety.html [https://perma.cc/PA9C-HET6].

69. Brad Smith, A Tragedy that Calls for More than Words: The Need for the Tech Sector to Learn and Act After Events in New Zealand, Microsoft (Mar. 24, 2019), https://blogs.microsoft.com/on-the-issues/2019/03/24/a-tragedy-that-calls-for-more-than-words-the-need-for-the-tech-sector-to-learn-and-act-after-events-in-new-zealand/ [https://perma.cc/ML64-JQTF].

Mark Zuckerberg concisely explained Facebook's shift in thinking this way: "It's not enough to just connect people. We have to make sure that those connections are positive. It's not enough to just give people a voice. We need to make sure that people aren't using it to harm other people or to spread misinformation."[70]

There are many more examples, but the point is that the major platforms now aspire to effective speech control to protect the "health" or "safety" of their platforms. But how to they do it? That is the subject of the next section.

B. *How Platforms Control Speech*

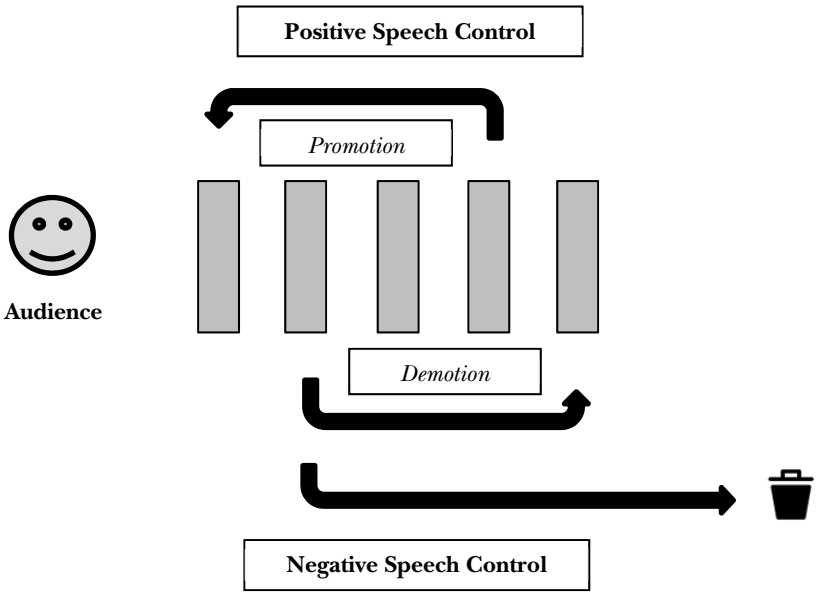
The control of speech in the United States and the world is possibly the most advanced example of a hybrid human–machine system of social ordering that has replaced what was once primarily governed by law. All of the major speech platforms use a mixture of software, humans following rules, and humans deliberating to enforce and improve their content rules.[71]

---

70. Transcript of Mark Zuckerberg's Senate Hearing, Wash. Post (Apr. 10, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/ (on file with the *Columbia Law Review*).

71. For recent articles offering a deeper investigation into how these platforms are shaping their content rules, see Kate Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, 131 Harv. L. Rev. 1598 (2018), and Simon van Zuylen-Wood, "Men Are Scum": Inside Facebook's War on Hate Speech, Vanity Fair (Feb. 26, 2019), https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech [https://perma.cc/3D7P-773L]. The author also attended a two-day Facebook seminar on its speech-control systems, from which some of this information is drawn.

FIGURE 1: HOW PLATFORMS CONTROL SPEECH



Speech is controlled by both affirmative and negative tools (promotion and suppression). *Affirmative* speech control entails choosing what is brought to the attention of the user. It is found in the operation of search results, newsfeeds, advertisements, and other forms of promotion and is typically algorithmic.[72] *Negative* speech control consists of removing and taking down disfavored, illegal, or banned content, and punishing or removing users.[73] The latter form of control, inherently more controversial, may be achieved in response to complaints, or proactively, by screening posted content.

Both positive and negative speech control have both human and algorithmic elements. Google's search results, the Facebook newsfeed, and the order in which tweets appear to Twitter users are all decided by algorithm.[74] In recent years, platforms like Facebook and Twitter have

72. See, e.g., How Search Algorithms Work, Google, https://www.google.com/search/howsearchworks/algorithms/ [https://perma.cc/RY8V-58ZQ] [hereinafter Search Algorithms] (last visited July 31, 2019).

73. Miguel Helft, Facebook Wrestles with Free Speech and Civility, N.Y. Times (Dec. 12, 2010), https://www.nytimes.com/2010/12/13/technology/13facebook.html?_r=0 (on file with the *Columbia Law Review*).

74. See, e.g., Search Algorithms, supra note 72; Nicolas Koumchatzky & Anton Andryeyev, Using Deep Learning at Scale in Twitter's Timelines, Twitter: Blog (May 9, 2017), https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html [https://perma.cc/E464-DJKU]; Ramya Sethuraman, Jordi Vallmitjana & Jon Levin, Using Surveys to Make News Feed More Personal, Facebook

begun using their affirmative powers of promotion to disadvantage disfavored speech, by ranking it as "lower quality."[75] For example, facing criticism that it had aided the dissemination of fake news and propaganda, Facebook in 2017 announced it was reworking its algorithm to disfavor, among other things, posts that were untruthful.[76] And as part of its technical attack on abusive speech, Twitter suggested that its search results would avoid content algorithmically deemed abusive or of low quality.[77]

The negative methods of speech control on platforms—takedowns—were originally complaint driven and executed by humans.[78] On the major platforms, the takedowns were first implemented for nudity and pornography. Platforms like Facebook and YouTube kept pornography off of their platforms by employing humans to swiftly respond to complaints and took down almost all nudity or pornographic films.[79] Today, those systems have matured into large "content review systems" that combine human and machine elements.

Facebook has been the most transparent about its system. The human part is some 15,000 reviewers, most of whom are private contractors working at call centers around the world, coupled with a team of technical and legal experts based in Facebook's headquarters.[80]

Newsroom (May 16, 2019), https://newsroom.fb.com/news/2019/05/more-personalized-experiences/ [https://perma.cc/U9JL-QVUV].

75. See, e.g., Varun Kacholia, News Feed FYI: Showing More High Quality Content, Facebook Bus. (Aug. 23, 2013), https://www.facebook.com/business/news/News-Feed-FYI-Showing-More-High-Quality-Content [https://perma.cc/422G-Z4UJ] (stating that Facebook's machine-learning algorithm counts reports that a post is "low quality" in deciding what content to show).

76. Adam Mosseri, Working to Stop Misinformation and False News, Facebook for Media (Apr. 7, 2017), https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news [https://perma.cc/7D9L-8GKQ].

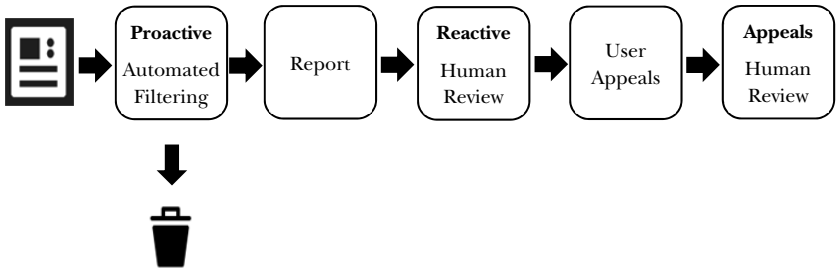
77. See Donald Hicks & David Gasca, A Healthier Twitter: Progress and More to Do, Twitter: Blog (Apr. 16, 2019), https://blog.twitter.com/en_us/topics/company/2019/health-update.html [https://perma.cc/SRF7-Q2UR].

78. See id. (stating that Twitter previously relied on reports to find abusive tweets).

79. See Nick Summers, Facebook's 'Porn Cops' Are Key to Its Growth, Newsweek (Apr. 30, 2009), https://www.newsweek.com/facebooks-porn-cops-are-key-its-growth-77055 [https://perma.cc/5HRA-UMMM] (describing the job of Facebook's content moderators and the scope of its review system); Catherine Buni & Soraya Chemaly, The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech, The Verge (Apr. 13, 2016), https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-YouTube-facebook-reddit-censorship-free-speech [https://perma.cc/U8PS-H6J8] (describing the job of content moderators in reviewing posts); Jason Koebler & Joseph Cox, The Impossible Job: Inside Facebook's Struggle to Moderate Two Million People, Vice (Aug. 23, 2018), https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works [https://perma.cc/4VFY-5FZ5] (describing the history of Facebook's content moderation system).

80. See van Zuylen-Wood, supra note 71; Casey Newton, The Trauma Floor: The Secret Lives of Facebook Moderators in America, The Verge (Feb. 25, 2019), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona/ [https://perma.cc/F78T-AJY3].

FIGURE 2: CONTENT REVIEW AT FACEBOOK



In this system, forbidden content is flagged, and then sent to a human for review. If the human decides it violates the content guidelines, they take it down, and a notice is sent to the poster, who may ask for an appeal. The appeal is decided by a human; in hard cases, the appeal may go through several levels of review.[81]

In recent years, Facebook and the rest of the platforms have deployed intelligent software as an aid to this process. The first lines of defense are proactive filters which prevent certain forms of content from being posted at all. Among the first AI-driven negative speech controls was YouTube's Content ID system, first launched in 2007.[82] Content ID is software that compares uploaded videos against a database of copyrighted materials to determine whether the video is presumptively infringing a copyright.[83] If so, the copyright owner is automatically notified and given the choice of ordering the video taken down, or accepting a revenue-sharing agreement for any advertising revenue the video generates.[84] Since 2013 or so, the major platforms have used a similar system, PhotoDNA, that proactively detects videos of child pornography

81. See van Zuylen-Wood, supra note 71 (describing the development of Facebook's appeals process).

82. See Google, How Google Fights Piracy 24 (2018), https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Piracy_2018.pdf [https://perma.cc/9VHW-NX35] [hereinafter How Google Fights Piracy]; see also Sam Gutelle, The Long, Checkered History of YouTube's Attempt to Launch a Music Service, Tubefilter, (May 22, 2018), https://www.tubefilter.com/2018/05/22/youtube-music-service-history/ [https://perma.cc/D82U-P6HC] (describing the "wild west" history of the early days of music video distribution on YouTube prior to the launch of Content ID).

83. See YouTube Help, How Content ID Works, supra note 39.

84. See id. According to Google, the arrangement has yielded payments of over $3 billion for rights holders. How Google Fights Piracy, supra note 82, at 25.

and prevents them from being posted.[85] The major platforms have also installed proactive screens to block terrorist propaganda.[86] Hence, in testimony before Congress, Zuckerberg stated that "99 percent of the ISIS and Al Qaida content that we take down on Facebook, our A.I. systems flag before any human sees it."[87]

Proactively flagging hate speech and other forms of offensive speech is inherently more subjective than flagging nudity, copyright infringement, or child pornography. Nonetheless, Twitter and Facebook have begun using software to flag or take down such materials.[88] At Twitter in 2017, Dorsey pledged "a completely new approach to abuse" involving more proactive use of AI.[89] Twitter redesigned its search engine to create the option of hiding abusive content;[90] the platform also began systematically downgrading "low-quality" tweets.[91]

But what about the hard cases? In 2018, Facebook announced that it was planning to supplement its current review process with review conducted by a review board, acting in a court-like fashion, comprised of

---

85. See Riva Richmond, Facebook's New Way to Combat Child Pornography, N.Y. Times: Gadgetwise (May 19, 2011), https://gadgetwise.blogs.nytimes.com/2011/05/19/facebook-to-combat-child-porn-using-microsofts-technology/ (on file with the *Columbia Law Review*) (reporting Facebook's adoption of PhotoDNA technology); Jennifer Langston, How PhotoDNA for Video Is Being Used to Fight Online Child Exploitation, Microsoft (Sept. 12, 2018), https://news.microsoft.com/on-the-issues/2018/09/12/how-photodna-for-video-is-being-used-to-fight-online-child-exploitation/ [https://perma.cc/2LKS-RBMN].

86. See Klonick, supra note 71, at 1651–52 (describing how Facebook, YouTube, and Twitter came to monitor and remove terrorist content at the request of the government, then later on their own); Joseph Menn & Dustin Volz, Google, Facebook Quietly Move Toward Automatic Blocking of Extremist Videos, Reuters (June 24, 2016), https://www.reuters.com/article/internet-extremism-video/rpt-google-facebook-quietly-move-toward-automatic-blocking-of-extremist-videos-idUSL1N19H00I [https://perma.cc/25XD-9D9N].

87. Transcript of Mark Zuckerberg's Senate Hearing, supra note 70.

88. See Daniel Terdiman, Here's How Facebook Uses AI to Detect Many Kinds of Bad Content, Fast Company (May 2, 2018), https://www.fastcompany.com/40566786/heres-how-facebook-uses-ai-to-detect-many-kinds-of-bad-content [https://perma.cc/N924-NACX] (reporting on the details of Facebook's AI content-flagging system); Queenie Wong, Twitter Gets More Proactive About Combating Abuse, CNET (Apr. 17, 2019), https://www.cnet.com/news/twitter-gets-more-proactive-about-combating-abuse/ (on file with the *Columbia Law Review*) (noting Twitter claims that thirty-eight percent of all content that violates its terms of service is flagged automatically before a user reports it).

89. Jack Dorsey (@jack), Twitter (Jan. 30, 2017), https://twitter.com/jack/status/826231794815037442 [https://perma.cc/7HGX-YV49]; see also Kurt Wagner, Twitter Says It's Going to Start Pushing More Abusive Tweets Out of Sight, Vox: Recode (Feb. 7, 2017), https://www.vox.com/2017/2/7/14528084/twitter-abuse-safety-features-update [https://perma.cc/6GQD-JBVA].

90. Wagner, supra note 89 (reporting on Twitter's "safe search" feature and its use of "machine learning technology (a.k.a. algorithms) to automatically hide certain responses," which the user cannot opt out of).

91. Jane Wakefield, Twitter Rolls Out New Anti-Abuse Tools, BBC (Feb. 7, 2017), https://www.bbc.com/news/technology-38897393 [https://perma.cc/2TC2-7MS4].

external, disinterested parties.[92] It is to that and other adjudicative bodies to which we now turn.

C. *The Reinvention of Adjudication*

Speech control by its nature produces hard problems. Is the phrase "kill all men" a form of hate speech, a joke (in context), or a feminist term of art?[93] If the phrase is taken down as hate speech, should it be put back up on review? The answer, of course, is that "it depends." The emergence of such problems has driven the major platforms to develop one or another forms of adjudication for this kind of hard case—a reinvention of the court, so to speak.

Lon Fuller defined an adjudication as a decision in which the participant is offered the opportunity to put forth "reasoned arguments for a decision in his favor."[94] By that measure, we can date the history of adjudicative content control on major online platforms to at least the mid-2000s.[95] In 2008, Jeffrey Rosen documented an early content-related deliberation at Google. It centered on a demand from the Turkish government that YouTube remove videos that the government deemed offensive to the founder of modern Turkey, in violation of local law. Rosen described the deliberation as follows:

> [Nicole] Wong [a Google attorney] and her colleagues set out to determine which [videos] were, in fact, illegal in Turkey; which violated YouTube's terms of service prohibiting hate speech but allowing political speech; and which constituted expression that Google and YouTube would try to protect. There was a vigorous internal debate among Wong and her colleagues at the top of Google's legal pyramid. Andrew McLaughlin, Google's director of global public policy, took an aggressive civil-libertarian position, arguing that the company should pro-

---

92. See Mark Zuckerberg, A Blueprint for Content Governance and Enforcement, Facebook (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/XU5H-Z2UV]; see also Ezra Klein, Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next, Vox (Apr. 2, 2018), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge [https://perma.cc/3N7X-AT2X] (noting Zuckerberg's early 2018 intention to form an independent appeal process for users to challenge Facebook's content-moderation decisions).

93. A description of the debate over this phrase can be found at stavvers, Kill All Men, Another Angry Woman (May 7, 2013), https://anotherangrywoman.com/2013/05/07/kill-all-men/ [https://perma.cc/59ES-5EEU].

94. Fuller, supra note 10, at 364.

95. There were also earlier such speech controls on online platforms. For a very early case study of an adjudication and punishment in an online forum, see Julian Dibbell, A Rape in Cyberspace, Village Voice (Oct. 18, 2005), https://www.villagevoice.com/2005/10/18/a-rape-in-cyberspace/ [https://perma.cc/2F5R-JVHY], See also Klonick, supra note 71, at 1618–21 (summarizing YouTube's, Facebook's, and Twitter's differing early approaches to content moderation, all overseen by lawyers normatively influenced by First Amendment principles).

> tect as much speech as possible. Kent Walker, Google's general counsel, took a more pragmatic approach, expressing concern for the safety of the dozen or so employees at Google's Turkish office. The responsibility for balancing these and other competing concerns about the controversial content fell to Wong, whose colleagues jokingly call her "the Decider . . . ."[96]

Since the mid-2010s, the platforms have developed larger and more specialized teams to adjudicate these kinds of hard problems, usually associated with the general counsel's office, and labeled the "Trust and Safety Council" (Twitter) or "safety and security" (Facebook).[97] Twitter, like Facebook, faces questions that emerge from complaints of abuse and propagandizing by political figures.[98] Its mechanism is a highly deliberative policy group centered in the general counsel's office to address the hardest speech problems.[99] Within the policy group is a leadership council that constantly updates the content guidelines applied by its reviewers.[100] The leadership council, which includes CEO Jack Dorsey, acts, in effect, as Twitter's supreme speech-moderation authority, and is responsible for both tough cases and large changes in policy.[101] It was through the deliberations of this group that, for example, Twitter decided to create more tools for screening "dehumanizing speech" in September 2018.[102] Here is how Twitter described its "dehumanizing speech" policy, outlined in a document not unlike that of a government agency promulgating a new rule:

> Language that makes someone less than human can have repercussions off the service, including normalizing serious violence. Some of this content falls within our hateful conduct policy . . . but there are still Tweets many people consider to be abusive, even when they do not break our rules. Better addressing this gap is part of our work to serve a healthy public conversation.

---

96. Jeffrey Rosen, Google's Gatekeepers, N.Y. Times Mag. (Nov. 28, 2008), https://www.nytimes.com/2008/11/30/magazine/30google-t.html (on file with the *Columbia Law Review*).

97. See Sara Harrison, Twitter and Instagram Unveil New Ways to Combat Hate—Again, WIRED (July 11, 2019), https://www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again (on file with the *Columbia Law Review*).

98. See Interview by Will Oremus with Vijaya Gadde, Gen. Counsel, Twitter (July 19, 2018), https://slate.com/technology/2018/07/twitters-vijaya-gadde-on-its-approach-to-free-speech-and-why-it-hasnt-banned-alex-jones.html [https://perma.cc/VPY8-SWM2] (quoting Twitter's general counsel as saying that "philosophically, [Twitter] ha[s] thought very hard about how to approach misinformation, and . . . felt that we should not as a company be in the position of verifying truth").

99. See id.

100. Telephone Interview with Vijaya Gadde, Legal, Policy, & Trust and Safety Lead, Twitter (Mar. 29, 2019).

101. Id.

102. See Vijaya Gadde & Del Harvey, Creating New Policies Together, Twitter: Blog (Sept. 25, 2018), https://blog.twitter.com/en_us/topics/company/2018/Creating-new-policies-together.html [https://perma.cc/W6TR-EJS9].

> With this change, we want to expand our hateful conduct policy to include content that dehumanizes others based on their membership in an identifiable group, even when the material does not include a direct target.[103]

We have already discussed Facebook's basic system of review.[104] Similar to Twitter, it currently has an internal policy group that works on hard cases and updates to policies in response to such cases.[105] To supplement and replace parts of the appeal process, the firm in 2018 announced plans to create an independent review board.[106] As Zuckerberg explained the idea,

> You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don't work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of people all around the world.[107]

According to Facebook, the board would be independent, with approximately forty members, and sit in panels of three[108] to review "hard cases."[109] They would be brought the hardest questions arising from content control on Facebook, and release their written decisions in two weeks.[110] The panels would have the ability to overrule Facebook's decisions and make policy suggestions, but not to rewrite the content rules themselves.[111]

Here, in summary form, we have a sense of how machines and humans combine to control speech on the major online platforms. We can now address the question of whether this institutional framework offers any promise for the future.

III. HYBRID SYSTEMS AND THE COMPARATIVE ADVANTAGES OF SOFTWARE AND COURTS

This Part addresses the comparative advantages of software and courts, and offers a normative defense of hybrid systems.

---

103. Id.

104. See supra notes 75–81 and accompanying text.

105. See Klein, supra note 92.

106. See supra note 92 and accompanying text.

107. Klein, supra note 92.

108. Facebook, Draft Charter: An Oversight Board for Content Decisions 3 (2019), https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-2.pdf [https://perma.cc/E9ZX-EDVF] [hereinafter Draft Charter].

109. Nick Clegg, Charting a Course for an Oversight Board for Content Decisions, Facebook Newsroom (Jan. 28, 2019), https://newsroom.fb.com/news/2019/01/oversight-board [https://perma.cc/3F6Q-EZ9X].

110. Draft Charter, supra note 108, at 5.

111. Id. at 3.

A. *Will Software Eat the Law?*

The case study of the online control of speech has shown the tendency of rule-based systems to generate hard and easy cases, giving rise to crude hybrid systems designed to manage that challenge. This Part seeks to theorize some of the advantages of hybrid systems. This returns us to the central question: Will software eat the law? (Or, as asked here, will software tools take over almost all of online content control?) In our case study, for routine matters, the answer is already yes, because of the undeniable comparative advantage of software in matters of scale, speed, and efficacy. To ask this question is a little like asking, a century ago, whether motorized lawnmowers might take over the mowing of lawns. But as to whether software will or should replace everything, the answer is no.

It is important to be more precise as to why this is so. As a means of regulation, software's main advantage over legal systems lies in what law would call its enforcement capacity.[112] Code is fast, can scale to meet the size of the problem, and operates at low marginal cost. But there is more to it than that. Code can be designed, as Lessig first pointed out, to change the very architecture of decision, the option set, and the menu of choices faced.[113] Consider that, when it comes to child pornography, the main platforms don't just ban it and punish transgressors but remove the option of posting it in the first place.[114] The enforcement mechanism is therefore ex ante rather than ex post, and hence vastly more effective than law, which always acts after a wrong is committed.

But if intelligent software is effective, it is also inherently inhuman, and prone, at least for the foreseeable future, to make absurd errors that can be funny, horrific, or both. Following rules blindly leads to consequences like the takedown of famous paintings as "nudity."[115] Software also faces limits of explainability, which is a problem for legal decision-making. Software can often explain *how* it reached a decision, but not *why*.[116] That may be fine for a thermostat, but is a limitation for a system that is supposed to both satisfy those subjected to it and prompt acceptance of an adverse ruling.

As it stands, the decisions to take down content by Facebook or Twitter are, to users, nearly a black box, which is acceptable for routine

---

112. See Wu & Talley, supra note 7.

113. See Lessig, supra note 12, at 121–25.

114. Richmond, supra note 85.

115. See, e.g., Kerry Allen, Facebook Bans Flemish Paintings Because of Nudity, BBC: News from Elsewhere (July 23, 2018), https://www.bbc.com/news/blogs-news-from-elsewhere-44925656 [https://perma.cc/JH5Z-CR9P].

116. See Ashley Deeks, The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. 1829, 1832–38 (2019); cf. Tom Simonite, Google's AI Guru Wants Computers to Think More Like Brains, WIRED (Dec. 12, 2018), https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/ [https://perma.cc/ECJ7-3P5Z].

decisions, but in borderline cases have already provoked anger and dissatisfaction.[117] As Richard M. Re and Alicia Solow-Niederman warn, software decision-systems can be "incomprehensible, data-based, alienating, and disillusioning."[118]

This is what the speech control case study helps make clear. If the only goal in speech control was taking down as much forbidden material as quickly as possible, mistakes be damned, the discussion would be over. But you don't need to be a First Amendment scholar to suggest that this would hardly amount to a satisfying or successful system of speech control, or one that the public would accept. The lines governing the forbidden from the provocative are always fuzzy, and building a healthy speech environment, at the risk of stating the obvious, is more than building the fastest takedown machine. In fact, the engineer's thirst for efficacy can obscure the fact that what Facebook and other platforms are building can also be described, without exaggeration, as among the most comprehensive censorship machines ever built.

Nor can we ignore the fact that what counts as acceptable speech for billions of people around the world is currently being decided by a relatively small group of private actors in Northern California. To suggest that this creates questions of legitimacy in the decision of matters of interest to the public in many countries seems almost too obvious to state. Hence, based on both public dissatisfaction *and* poor results, a purely software-based replacement is a bad aspiration.

That's why the platforms are already turning to specialized human adjudicators, as a supplement to the software systems, to offer answers to some of these problems.[119] Their advantages—really the advantages of courts more generally—lie in two areas.

The first is procedural fairness. A group of legal theorists, of which Tom Tyler is best known, has for decades suggested that the best justification for the court system lies in providing a sense of procedural fairness to participants.[120] The empirical studies conducted by Tyler and others suggest that when litigants feel they have a voice and are treated with respect, they tend to be more accepting of decisions, even adverse outcomes.[121] It seems unlikely, in the near future, that people with a grievance will be more satisfied with a software decision than a human decision on an important complaint. In the future, having a major decision be made by a human may become a basic indicium of fairness; it

117. See, e.g., Sam Levin, Julia Carrie Wong & Luke Harding, Facebook Backs Down from 'Napalm Girl' Censorship and Reinstates Photo, Guardian (Sept. 9, 2016), https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo [https://perma.cc/78CB-KN32].

118. Re & Solow-Niederman, supra note 2, at 242.

119. See supra text accompanying notes 107–111.

120. See Tom R. Tyler, Procedural Justice and the Courts, 44 Ct. Rev. 26, 30–31 (2007) (summarizing research in this area).

121. See id. at 26.

is implicit in the emergence of what Aziz Z. Huq calls a "right to a human decision."[122]

That said, it is possible that our taste for human adjudication might be fleeting; perhaps it is akin to an old-fashioned taste for human travel agents. Eugene Volokh argues that any preference for human decision may turn out to be temporary, because humans are imperfect as well.[123] He believes that if an AI judge produces good decisions and good opinions, it will be broadly accepted, particularly if it is cheaper for users.[124] Volokh, characteristically, overstates his point, but he is right that there are in fact many areas where "impartial" code is trusted more than humans (compare Google Maps to asking for directions).[125] But that acceptance turns very heavily on the quality of decisions, to which we now turn.

The second benefit of human courts over software is their advantages in hard cases, and the prevention of absurd errors, obviously unjust results, and other inequitable consequences of a blind adherence to rules. There are, on closer examination, several ways in which a case can be "hard." Some cases might be hard only because the software lacks the ability to understand context or nuance, as in understanding that "I'm going to kill my husband" may be a figurative statement, not a death threat. And, others may be hard in the jurisprudential sense because they require the balancing of conflicting values or avoidance of absurd consequence. Finally, it may be that the stakes just seem large enough to merit human involvement, as in the decision to sentence someone to death. In each of these cases, the use of humans may prevent what Re and Solow-Niederman believe will be a tendency of AI systems to promote "codified justice at the expense of equitable justice."[126] How so? The premise is that leaving the hard cases to people of good character who are asked to listen to reasoned argument will have an effect, and that the effect will be positive for the rule system in question.

The theoretical support for this position is one of ancient pedigree and comes from the idea that something happens when intelligent, experienced, and thoughtful humans are asked to hear reasoned argument and the presentation of proofs to determine how a dispute should be settled. Over the centuries, the mental process accompanying the judicial process has been described in many different ways.[127] In the Anglo American tradition, it was presented in the semi-mystical idea that judges "discover" the law in the process of adjudication and deliberation, a law

---

122. Huq, supra note 14, at 2.

123. See Volokh, supra note 2, at 1170–71.

124. Id.

125. See also Tim Wu, The Bitcoin Boom: In Code We Trust, N.Y. Times (Dec. 18, 2017), https://www.nytimes.com/2017/12/18/opinion/bitcoin-boom-technology-trust.html (on file with the *Columbia Law Review*) (arguing that there are sometimes reasons to trust in code).

126. Re & Solow-Niederman, supra note 2, at 255, 260.

127. See id. at 252–53.

that was usually thought to be God given.[128] Blackstone writes of judges discovering the "the eternal, immutable laws of good and evil, to which the creator himself in all his dispensations conforms; and which he has enabled human reason to discover, so far as they are necessary for the conduct of human actions."[129]

Blackstone's theory that the law is best discovered by tuning into heavenly emanations enjoys a more limited following today.[130] But the idea that a particular mental process accompanies adjudication survives, even in the work of those highly critical of natural law reasoning. It is found in Llewellyn's idea of a judge's understanding of the "real rules" as distinct from the paper rules, and the skill involved in weighing demands of flexibility and stability in a legal system.[131] The judicial process is also a major part of Ronald Dworkin's theory of legal reasoning, which suggests that judges, when facing hard cases, begin to fill in gaps or conflicts through a process of rights-driven moral reasoning.[132] Hence, as Dworkin wrote in *Taking Rights Seriously*, a court won't let the son who murders his grandfather inherit wealth, not based on the following of any rule, but by reaching for the principle that doing so would be morally wrong.[133]

One does not need not to accept or agree with Dworkin's particular theory of how judges decide hard cases to accept that he has gotten at something important in the mechanism of decisionmaking. Richard Posner, for example accepts the premise that a judge, when deciding a hard case, exercises powers of intuitive judgment, though Posner believes they should be powers of pragmatic judgment.[134] Posner, who was a judge, wrote of the process this way: Judges necessarily "consider the implication of [their] interpretation for the public good" and, when making decisions about private rights, "consider the social consequences of alternative answers."[135] Or perhaps Fuller was correct when he asserted that the key is not labeling a person a judge, so much as the entire process of adjudication. He located the special sauce, such as it is, as "the presentation of proofs and reasoned arguments," yielding an expectation that the decision "meet the test of rationality."[136]

Cynics reading the preceding paragraphs might think that all that is being described is a bunch of hoodoo voodoo, a mystic secret sauce that

---

128. 1 William Blackstone, Commentaries *40.

129. Id.

130. See, e.g., John S. Baker, Jr., Natural Law and Justice Thomas, 12 Regent U. L. Rev. 471, 471 (1999) (describing and defending the use of natural law approaches).

131. See Frederick Schauer, Introduction to Llewellyn, *in* The Theory of Rules, supra note 17, at 11–13.

132. See Dworkin, supra note 16, at 81–88.

133. Id. at 23–28.

134. See Richard A. Posner, Pragmatic Adjudication, 18 Cardozo L. Rev. 1, 5–8 (1996).

135. Richard A. Posner, What Am I? A Potted Plant?: The Case Against Strict Constructionism, New Republic, Sept. 28, 1987, at 23, 23.

136. See Fuller, supra note 10, at 365–70.

is hiding nothing more than judicial whim. Be that as it may, even such whims remain hard to replicate using artificial intelligence. And what Blackstone, Llewellyn, Dworkin, and Posner are all getting at is familiar to anyone who has either sat as a judge, or been asked to decide a hard case. The process brings forth a series of instincts, competing intuitions that can be of differing strengths in different people, but whose existence cannot be denied. A good account is given by Benjamin Cardozo, who, in *The Judicial Process*, describes a judge in deliberation as bombarded by competing forces, not all conscious.[137] Judges often ruminate at length, change their mind, want more facts, and want to consider different futures based on what they are proposing to do. Some may secretly (or openly, like Blackstone) believe that they are tapping into the divine, or, for those who claim a more secular mindset, the immutable principles of moral philosophy.

That said, returning to this Essay's case study and our times, it must be admitted that hoping for a Herculean process of judicial reasoning may be expecting a lot from the first hybrid systems, like the Facebook review board and its part-time judges. The court will have many disadvantages, including a lack of history, lack of traditions, lack of connection with government, and smaller matters like the probable lack of a courtroom (though perhaps robes will be provided). Fuller's idea that the setting and context matter may be right, and if so the Facebook appeals board may never inspire the kind of reasoning that garners respect.

In contrast, while I doubt it, it is possible that AI systems will soon begin to replicate the adjudicatory function in a manner indistinguishable from a human, while becoming able to explain what they are doing in a manner that complainants find acceptable.[138] And that, perhaps, will inspire people to trust such programs as less fallible than humans. Then the question will be whether judges are more like travel agents or more like spouses—whether being human is essential to the role. But for the foreseeable future, there is nothing that has anything close to these abilities; what we have is software intelligent enough to follow rules and replicate existing patterns, but that's about it. That's what makes hybrid systems seem almost inevitable, at least should we want social ordering to have any regard for the demands of justice, equity, or other human values.

B. *Implications and Other Counterarguments*

Reflecting their roots as software companies, the leaders of Silicon Valley firms usually state their ambition to have intelligent software even-

137. See Cardozo, supra note 18, at 10–12.

138. See Louise A. Dennis & Michael Fisher, Practical Challenges in Explicit Ethical Machine Reasoning, ArXiv (Jan. 4, 2018), https://arxiv.org/pdf/1801.01422.pdf [https://perma.cc/69UG-G3FK] (reviewing the several practical challenges AI systems face in replicating ethical reasoning).

tually solve problems by replacing humans entirely.[139] For example, self-driving cars are not designed as aids to driving, but as replacements for human drivers.[140] The industry has expressed similar goals for content-control systems, as in Facebook's promise to Congress that control of hate speech will be automated in the next five to ten years.[141]

This is the wrong aspiration. While the desire to have more effective and efficient systems of social control is understandable, far too much would be lost. The implication of this Essay is that the designers of intelligent software produced for social ordering should be aiming for the autopilot, not the self-driving car. The reasons why have already been stated; but until a computer is able to replicate not only a judge but the entire process of adjudication, we remain far short of an AI solution.

Similarly, as government begins to automate parts of the legal system (as has happened in limited ways already), a hybrid system should be the aspiration as well. Routine matters, like routine motion practice, and even perhaps frivolous cases, might well be automated to reduce the workload of the judiciary. The promise of doing so is not just saving costs but giving the judiciary more room to emphasize justice in the individual case as it devotes less of its time to reducing the judicial workload. Since the 1980s, numerous critics have pointed out that the huge increases in federal court filings have created a workload crisis.[142] As Judge Roger Miner wrote in 1997, "The situation has been deteriorating for many years and, although the courts have been attempting to cope by using various methods to accommodate the growing caseload traffic, the problems associated with volume largely remain unresolved."[143] One reaction has been the creation of various judicial doctrines designed to cope with the workload, from easier standards of dismissals, various means of reducing jurisdiction, plea bargaining in criminal cases, and reduced oral arguments.[144] With the help of software to handle routine procedural matters and even the decision of routine motions, government courts and judges might be able to devote more time and effort to the hard cases and im-

---

139. See, e.g., Kevin Roose, A Machine May Not Take Your Job, but One Could Become Your Boss, N.Y. Times: The Shift (June 23, 2019), https://www.nytimes.com/2019/06/23/technology/artificial-intelligence-ai-workplace.html (on file with the *Columbia Law Review*).

140. See Jonathan Vanian, Will Replacing Human Drivers with Self-Driving Cars Be Safer?, Fortune (June 14, 2017), http://fortune.com/2017/06/14/ford-argo-ai-self-driving-cars/ [https://perma.cc/7KWH-YAGU] ("[A]ccording to Bryan Salesky, the CEO of [Ford Motor Company's subsidiary] Argo AI, . . . [t]he rise of self-driving cars will usher a 'much safer mode of transportation' by 'removing the human from the loop' . . . .").

141. Transcript of Mark Zuckerberg's Senate Hearing, supra note 70.

142. See Cara Bayles, Crisis to Catastrophe: As Judicial Ranks Stagnate, 'Desperation' Hits the Bench, Law360 (Mar. 19, 2019), https://www.law360.com/articles/1140100 [https://perma.cc/YW3D-3ULZ]; see also Roger J. Miner, Book Review, 46 Cath. U. L. Rev. 1189, 1189–91 (1997) (reviewing Richard A. Posner, The Federal Courts: Challenge and Reform (1996) [hereinafter Posner, The Federal Courts]).

143. Miner, supra note 142, at 1189.

144. See Posner, The Federal Courts, supra note 142, at 160–85.

provement of the rules without the need to be constantly concerned about the impact of their decisions on their own workload.

This Essay could be wrong either on descriptive or normative grounds. Descriptively, it could turn out to be wrong that human judges have any lasting advantages over software; if an AI can pass a Turing test, it might well soon begin to replicate that which we call justice, and people could get used to decisions made by a machine. Or the opposite could be true: AI has often been grossly overrated and software might not make the inroads expected, leaving the legal system and other systems of social ordering more or less intact. There is no real way to address either of these objections other than to say that prediction is hard, especially when it comes to the future.

Normatively, it could also be wrong to think that there is really anything appealing about a hybrid human–machine system. Anthropologists like Hugh Gusterson write about the rise of the "roboprocess"—systems, like the U.S. credit rating system, that combine software with humans but actually disempower and deskill the humans employed by them.[145] Re and Solow-Niederman argue that introducing more software into the justice system will drive a shift in norms toward "codified" (that is, rule-driven) justice, as opposed to equitable justice.[146] They are not optimistic about adding humans, believing that "[r]etaining a human in the system . . . could succeed in preserving the legal system's preexisting public legitimacy—but only by objectionably sacrificing efficiency and uniformity that pure AI adjudication would otherwise offer."[147] The worst version of the hybrid system would pair the unthinking brutality of software-based justice with a token human presence designed to appease the humans subject to it. Such a system might arise out of cost cutting, in the manner that automated assistants are used in customer support to save money rather than improve service. This argument does make clear the danger of judging the judicial system by its costs alone, when the stakes are so much higher.

This suggests that the key question is the human–machine interface in a hybrid system. Just when and why are decisions brought to human attention, and who decides when a human should decide? Stated differently, how do we distinguish between "easy" and "hard" questions? It quickly becomes apparent that the human cases must include not just those that are hard in a jurisprudential sense, but also those where the stakes are large. The automated dispenser of speeding tickets may be one thing, but it is hard to imagine the fully automated assignment of the

145. See Hugh Gusterson, Introduction: Robohumans, *in* Life by Algorithms, How Roboprocesses Are Remaking Our World, 1, 13–26 (Catherine Besteman & Hugh Gusterson, eds. 2019).

146. See Re & Solow-Niederman, supra note 2, at 246–47.

147. Id. at 284–85.

death sentence, even if it were shown, as compared to a jury, to more reliably determine guilt or innocence.[148]

Deciding what and when questions go to an empowered human is difficult out of context, but the most obvious model is a certiorari system used by appellate courts—a human system designed to decide when to decide. Whether that system is itself human or machine-run, or another hybrid makes for an interesting design problem. In any event, setting the border between human and machine decision is surely the linchpin of a successful hybrid system.

It might also be that hybrid systems accelerate a privatization of public justice. For some decades, with the rise of measures like compulsory arbitration, critics have complained that American justice has been privatized, usually in a manner designed to disfavor consumers, patients, and other weak groups.[149] The hybrid systems in the case study are all private adjudicators and policymakers. Their speech codes are created in-house, without traditional forms of public input. If successful, they may become a model whereby more and more areas of social ordering become subjects of such private hybrid systems.

It would be foolish to ignore such concerns. The topic of speech control may be a special case, given that the Supreme Court has effectively privatized speech control with its aggressive interpretations of the First Amendment.[150] But if we consider privatization of justice, the right answer might be "if you can't beat 'em, join 'em": The increased use of software may help improve the efficiency of routine justice, protecting the resources of the court system for preventing error in cases of either greater consequence or greater difficulty. Robot courts are not the right aspiration, but an augmented equivalent may very well be.

CONCLUSION

The comparative advantages of human, machine, and cyborg systems have been a longstanding subject of science fiction. But as the science fiction slowly becomes reality, one of the genre's longstanding predictions is coming true. It takes great effort to preserve human values when new technologies make it so easy to maximize efficient operations. There are reasons beyond the literary that so much science fiction is dystopian.

---

148. When it comes to war, a parallel debate concerns the deployment of autonomous weapons. See generally Amanda Sharkey, Autonomous Weapons Systems, Killer Robots and Human Dignity, 21 Ethics & Info. Tech. 75 (2019) (exploring criticisms of autonomous weapon systems as violating human dignity).

149. See, e.g., Jessica Silver-Greenberg & Michael Corkery, In Arbitration, a 'Privatization of the Justice System,' N.Y. Times: Dealbook (Nov. 1, 2015), https://www.nytimes.com/2015/11/02/business/dealbook/in-arbitration-a-privatization-of-the-justice-system.html (on file with the *Columbia Law Review*).

150. See, e.g., Reno v. ACLU, 521 U.S. 844, 870 (1997) (striking down the Communications Decency Act and holding that the internet is due the highest level of First Amendment protection).