

# Fair Learning

Mark A. Lemley\* & Bryan Casey\*\*

## Introduction

The challenge handed to the musician was peculiar and daunting: Take a five second sample of a randomly selected song and, with just a moment's notice, transform it into a full-length piece composed in the style of a completely different artist.<sup>1</sup> On this occasion, the musician rose to the challenge with such aplomb that it took the Internet by storm, earning praise and recognition from media outlets across the globe.<sup>2</sup> A Mozart concerto

---

© 2020 Mark A. Lemley & Bryan Casey.

\* William H. Neukom Professor, Stanford Law School; partner, Durie Tangri LLP.

\*\* Fellow, Center for Automotive Research at Stanford (CARS). Thanks to Ed Cavazos, Bob Glushko, Paul Goldstein, James Grimmelman, Rose Hagan, Daniel Hemel, Mark McKenna, Tyler Ochoa, Lisa Ramsey, Matthew Sag, Lea Shaver, Ben Sobel, Xiyin Tang, Josh Tucker, Rebecca Tushnet, and participants at the Works in Progress-Intellectual Property conference and workshops at the Stanford Symbolic Systems Program and the Chicago IP Colloquium for comments on a prior draft.

1. See *infra* notes 2–5 and accompanying text.

2. See, e.g., Will Knight, *This AI-generated Musak Shows Us the Limit of Artificial Creativity*, MIT TECH. REV. (Apr. 26, 2019), <https://www.technologyreview.com/s/613430/this-ai-generated-musak-shows-us-the-limit-of-artificial-creativity/> (describing MuseNet's ability to deftly accomplish these unique musical pairings); Devin Coldewey, *MuseNet Generates Original Songs*

played in the style of Katy Perry?<sup>3</sup> No problem. Lady Gaga's Poker Face in the style of a recent Bollywood hit? Why not?<sup>4</sup> Seemingly no pairing of styles, no matter how clashing, proved too difficult.<sup>5</sup>

The artist capable of such a *tour de force*? MuseNet—the type of unique, futuristic-sounding name common of many contemporary artists.<sup>6</sup> But, on closer inspection, MuseNet's name wasn't a reference to the internet (or even

---

*in Seconds, from Bollywood to Bach (or both)*, <https://techcrunch.com/2019/04/25/musenet-generates-original-songs-in-seconds-from-bollywood-to-bach-or-both> (same); Jon Porter, *OpenAI's MuseNet Generates AI Music at the Push of a Button*, THE VERGE (Apr. 26, 2019), <https://www.theverge.com/2019/4/26/18517803/openai-musenet-artificial-intelligence-ai-music-generation-lady-gaga-harry-potter-mozart> (same).

3. See Porter, *supra* note 2 (describing this ability).

4. See *id.*

5. The artist couldn't take on *literally* every pairing, of course, but was nonetheless capable of a stunning diversity of combinations. See Christine Payne, *MuseNet*, OPEN AI BLOG (Apr. 25, 2019), <https://openai.com/blog/musenet/> (describing some of the limits of musical diversity inherent to MuseNet).

6. See Porter, *supra* note 2.

to the Terminator series' dystopian Skynet). Rather, MuseNet was a neural network. The musician, in other words, was a robot,<sup>7</sup> not a person.<sup>8</sup>

---

7. Here, and for the remainder of this piece, we use the term robot loosely. See Bryan Casey & Mark Lemley, *You Might be A Robot*, 105 CORNELL L. REV. 287 (2020) (explaining the difficulties of explicitly defining the term robot and opting to include various forms of artificial intelligence (AI) in the definition).

8. A number of scholars have begun to address the copyrightability of creative works made by machines, following the lead of Pam Samuelson, who did it 34(!) years ago. Pamela Samuelson, *The Future of Software Protection: Allocating Ownership Rights in Computer-Generated Works*, 47 U. PITT. L. REV. 1185 (1986). See, e.g., Clark D. Asay, *Independent Creation in a World of AI* (working paper 2020); Jane C. Ginsburg & Luke Ali Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L.J. 343 (2019); William Ralston, *Copyright in Computer-Composed Music: Hal Meets Handel*, 52 J. COPYRIGHT SOC'Y USA 281 (2005); James Grimmelmann, *There's No Such Thing as a Computer-Authored Work - and It's a Good Thing*, 39 COLUM. J. L. & ARTS 403 (2016); Bruce E. Boyden, *Emergent Works*, 39 COLUM. J.L. & ARTS 377, 378 (2015); Annemarie Bridy, *Coding Creativity: Copyright and the Artificially Intelligent Author*, 5 STAN. TECH. L. REV. (2012); Enrico Bonadio & Nicola Lucchi, *How Far Can Copyright Be Stretched? Framing the Debate on Whether New and Different Forms of Creativity Can be Protected* (working paper 2020); Shlomit Yanivsky-Ranid, *Generating Rembrandt: Artificial Intelligence, Copyright, and Accountability in the 3a Era – The Human-Like Authors Are Already Here – a New Model*, 2017 MICH. ST. L. REV. 659; Hon. Katherine B. Forrest, *Copyright Law and Artificial Intelligence: Emerging Issues*,—J. COPYRIGHT SOC'Y U.S.A. 355. A Chinese court held in 2020 that an AI-written article is protected by copyright (and owned by the company that owns the AI). *Chinese Court: AI-Written Article Is Protected*, JurisNotes IP News, Jan. 22, 2020.

Other scholars have focused not on how copyright law affects AI, but how AI can be used in implementing copyright regimes, particularly fair use. See, e.g., Peter K. Yu, *Can Algorithms*

Thanks to rapid advances in a subfield of computer science known as machine learning (ML),<sup>9</sup> feats of robot ingenuity like the one displayed by MuseNet have become regular fixtures of the news. But feats of ML prowess aren't limited to displays of creativity. There have also been similarly impressive advances in a host of other industry and social contexts, ranging from transportation,<sup>10</sup> to media curation,<sup>11</sup> to medical diagnostics,<sup>12</sup> to insurance risk mitigation.<sup>13</sup>

---

*Promote Fair Use?* (working paper 2020); Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019). Our focus here is different than either thread. We are interested, not in how copyright might apply to the outcome of an AI, but how it affects the process of training and using that AI.

9. Although the term was first coined in 1959, see Arthur L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers. I.*, COMPUTER GAMES I. 335–365 (1988) (doi:10.1007/978-1-4613-8716-9\_14), the field did not see a significant inflection point in progress until around 2010.

10. See, e.g., Winnie Hu, *Driverless Cars Arrive in New York City*, N.Y. TIMES (Aug. 6, 2019), <https://www.nytimes.com/2019/08/06/nyregion/driverless-cars-new-york-city.html>.

11. See, e.g., Mike Isaac, *In New Facebook Effort, Humans Will Help Curate Your News Stories*, (Aug. 20, 2019), <https://www.nytimes.com/2019/08/20/technology/facebook-news-humans.html> (describing how “Facebook has long relied on algorithms to select news stories for its users to see”).

12. See Nicola Davis, *AI Equal with Human Experts in Medical Diagnosis, Study Finds*, THE GUARDIAN (Sept. 24, 2019), <https://www.theguardian.com/technology/2019/sep/24/ai-equal-with-human-experts-in-medical-diagnosis-study-finds>.

13. See Jason Pontin, *How AI-Driven Insurance Could Reduce Gun Violence*, WIRED (Feb. 27, 2018), <https://www.wired.com/story/how-ai-driven-insurance-could-reduce-gun-violence/>.

The vast potential of ML systems is matched only by their appetite for data. To perform, they must first learn how—generally, through a process of trial and error of epic proportions. And in order to create the right conditions for this learning process, engineers must begin by collecting and compiling enormous databases of exemplary tasks for machines to practice on, known as “training sets.”<sup>14</sup>

Enter copyright law. Creating a training set of millions of examples almost always requires, first, copying many more millions of images, videos, audio, or text-based works. Those works are almost all copyrighted.<sup>15</sup> Were a human to copy the sheer volume of songs that MuseNet did they’d be looking at serious consequences. (Just ask the founders of Napster;<sup>16</sup> or Cox Cable, which was ordered to pay \$1 billion because it didn’t terminate some

---

14. For a more thorough explanation of the term, see TRAINING AND TEST SETS: SPLITTING DATA, GOOGLE, <https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data>. For this Article’s purposes, the term “training set” encompassed both training sets (subsets of datasets used to train ML models) and test sets (subsets of datasets used to test a trained ML model).

15. There are occasional exceptions. A weather prediction AI, for instance, might train only on factual data provided by the National Weather Service, and that data is not subject to copyright. But as we discuss below, most training data sets are built on copyrighted works.

16. Staff and Agencies, *Napster Loses Net Music Copyright Case*, THE GUARDIAN (Jul. 26, 2000), <https://www.theguardian.com/technology/2000/jul/27/copyright.news>.

of its users who shared copyrighted files.)<sup>17</sup> But thanks in large part to copyright's fair use doctrine, robots that do the same have traditionally been granted broad latitude. As recently as 2016, James Grimmelmann observed that "quietly, invisibly, almost by accident, copyright has concluded that reading by robots doesn't count."<sup>18</sup>

Today, however, that truism is no longer quite so true. After decades of allowing—or even just plain ignoring—machine copying, copyright owners and courts have begun to loudly and visibly push back against the copyright system's permissive attitude towards machine copying.<sup>19</sup> True, a years-long saga between Google and the Author's Guild (hereafter, the Google Books Cases) offers hope to robot readers.<sup>20</sup> But countervailing, even contradictory, moves in other decisions<sup>21</sup> have thrown the legality of machine copying into question. Complicating matters more, the nature of machine copying has also changed as the use of datasets has expanded from narrower "text data

---

17. See Chris Eggerstsen, *Labels & Publishers Win \$1 Billion Piracy Lawsuit Against Cox Communications*, BILLBOARD (Dec. 19, 2019), <https://www.billboard.com/articles/business/legal-and-management/8546842/cox-1-billion-piracy-lawsuit-labels-publishers>.

18. James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 658 (2016).

19. See *infra* notes 21–30 and accompanying text.

20. See Part II.A *infra* (discussing the Google Books line of cases).

21. See, e.g., *Fox News Network, LLC v. TVEyes, Inc.*, 883 F. 3d 169 (2nd Cir. 2018); *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 543–44 (S.D.N.Y. 2013).

202x]

Fair Learning clean 92620

107

mining” (TDM) systems that read existing content to more sophisticated systems like MuseNet that actively learn from it. Given the doctrinal uncertainty and the rapid development of ML technology, it is unclear whether machine copying will continue to be treated as fair use.

There are reasons to think courts in the future won’t be so sympathetic to machine copying. Fair use doctrine in the last quarter century has focused on the transformation of the copyrighted work. ML systems, however, rarely transform the databases they train on; they are using the entire database, and for a commercial purpose at that. Courts may view that as a kind of free riding they should prohibit, particularly when the companies training ML models tend to be giant multinationals and the owners of individual photographs and books are often small, sympathetic plaintiffs. And many of those plaintiffs may be motivated to sue, either by the extraordinary statutory damages copyright law offers them or because they don’t want their work used to train an AI that might someday replace their job or might use the data in undesirable ways.<sup>22</sup>

---

22. We may learn more about the future of the transformative use doctrine this year as the Supreme Court takes up *Google v. Oracle*. *Google LLC v. Oracle America, Inc.*, No.18-956 (*cert. granted* Nov. 15, 2019). Aside from copyright, the bulk collection of training data also implicates a host of other laws that could prove formidable barriers in the future. *See, e.g.*, Kashmir Hill and Aaron Krolik, *How Photos of Your Kids Are Powering Surveillance Technology*, N.Y. TIMES (Oct. 11, 2019), <https://www.nytimes.com>

Further, these uncertainties arrive at a time when copyright's attitude toward robotic readership is under increasing fire in a court of equal importance to those established by Article III<sup>23</sup>—the court of public opinion.<sup>24</sup> Hardly a week now passes without headlines from media outlets, thought leaders, or advocacy organizations decrying new ML systems that push data usage norms to the limits.<sup>25</sup> Once a relatively obscure topic, debates

---

/interactive/2019/10/11/technology/flickr-facial-recognition.html(discussing how Illinois's Biometric Information Privacy Act of 2008 could have massive liabilities implications for companies using the photos of Illinois inhabitants); Orin Kerr, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143 (2016) (discussing how the Criminal Fraud and Abuse Act creates a chilling effect for those deploying the tools necessary to collect training data).

23. See U.S. CONST. art. III.

24. *Copyright Law Should Not Restrict AI Systems from Using Public Data* <https://www.datainnovation.org/2019/10/copyright-law-should-not-restrict-ai-systems-from-using-public-data/>.

25. See, e.g., Rachel Botsman, *Dawn of the Techlash*, GUARDIAN (Feb 10, 2018), <https://www.theguardian.com/commentisfree/2018/feb/11/dawn-of-the-techlash>; Eve Smith, *The Techlash Against Amazon, Facebook and Google—And What They Can Do*, ECONOMIST (Jan. 20, 2018), <https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do>; Gregory Bobillot, *'Techlash' – How Big Tech is Influencing Your Thinking*, FINANCIAL TIMES (Mat 10, 2018), <https://www.ft.com/video/3339f59e-f760-4bc7-b359-3899fabbd190>; *The Techlash Has Just Begun*, AXIOS (Jan. 10, 2018), <https://www.axios.com/the-techlash-1515609266-e27ca299-0031-460a-96f1-db842ec88121.html>; John Rubio, *'Tech Wreck,' 'Techlash,' 'Techmageddon' - Whatever You Call It, Wall Street Is Terrified Of It*, SEEKING ALPHA (Mar. 29, 2018), <https://seekingalpha.com/article/4159827-tech-wreck-techlash->



over ML and copyright law are now the subject of *New York Times* pieces headlined *How Photos of Your Kids Are Powering Surveillance Technology*<sup>26</sup> and takedown campaigns of “deepfake”<sup>27</sup> videos satirizing celebrities such as

---

techmageddon-whatever-call-wall-street-terrified; *Emails Show How Amazon is Selling Facial Recognition System to Law Enforcement*, ACLU OF NOR. CAL. (May 22, 2018), <https://www.aclunc.org/news/emails-show-how-amazon-selling-facial-recognition-system-law-enforcement>.

26. See, e.g., Kashmir Hill and Aaron Krolik, *How Photos of Your Kids Are Powering Surveillance Technology*, N.Y. TIMES (Oct. 11, 2019), <https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html> (noting: “Millions of Flickr images were sucked into a database called MegaFace. Now some of those faces may have the ability to sue.”); Olivia Solon, *Facial Recognition’s ‘Dirty Little Secret’: Millions of Online Photos Scraped Without Consent*, NBC NEWS (Mar. 17, 2019), <https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921>.

27. Bobby Chesney and Danielle Citron describe “deepfakes” as ML-based “[t]echnologies for altering images, video, or audio (or even creating them from scratch) in ways that are highly-realistic and difficult to detect. See Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security* 107 CAL. L. REV. 1753, 1757 (stating: “We use that label here more broadly, as shorthand for the full range of hyper-realistic digital falsification of images, video, and audio”). See also, e.g., Samantha Cole, *We Are Truly Fucked: Everyone is Making AI-Generated Fake Porn Now*, VICE: MOTHERBOARD (Jan. 24, 2018), [https://motherboard.vice.com/en\\_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley](https://motherboard.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley) (“[T]echnology[] allows anyone with sufficient raw footage to work with to convincingly place any face in any video.”).

Kim Kardashian.<sup>28</sup> The public reacted with fury when it learned that Clearview AI had collected over 3 billion photos from the internet to build a facial recognition database it licensed to law enforcement.<sup>29</sup> Prominent copyright professor Tim Wu tweeted that it “should be the target of a class action copyright lawsuit.”<sup>30</sup> This unique confluence of commercial, normative, and doctrinal factors has teed up what may well be one of the most important legal questions of the coming century: Will copyright law allow robots to learn?

In this Article, we argue that ML systems should generally<sup>31</sup> be able to use databases for training, whether or not the contents of that database are copyrighted. There are good policy reasons to do so. First, we need to

---

28. As Bobby Chesney and Danielle Citron note, “Pornographers have been early adopters of the technology, interposing the faces of celebrities into sex videos.” See Chesney & Citron, *supra* note 27. See also *id.* at 1793–1794 (floating copyright law as a potential, albeit imperfect, mechanisms for curbing deepfakes abuses); Alex Engler, *Fighting Deepfakes When Detection Fails*, Brookings (Nov. 14, 2019), <https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/> (same).

29. Beryl Lipton, *Records on Clearview AI Reveal New Info on Police Use*, <https://www.muckrock.com/news/archives/2020/jan/18/clearview-ai-facial-recognition-records/>.

30. <https://twitter.com/superwuster/status/1218524978225741824?lang=en>

31. The word “generally” here is intentional. As we discuss below, machine copying should not be permissible in every conceivable instance. Fair use in the machine learning context, for example, should be sensitive to the purpose of the ML system and what it eventually produces as output.

encourage people to compile new databases and to open them up for public scrutiny or innovation. Broad access to training sets will further these objectives, ultimately making artificial intelligence systems using ML algorithms better, safer, and fairer.<sup>32</sup> Second, an ML system’s use of the data often *is* transformative as that term has come to be understood in copyright law, because even though it doesn’t change the underlying work, it changes the purpose for which the work is used.<sup>33</sup> And because training sets are likely to contain millions of different works with thousands of different owners, there is no plausible option simply to license all of the underlying photographs, videos, audio files, or texts for the new use. So allowing a copyright claim is tantamount to saying, not that copyright owners will get paid, but that the use won’t be permitted at all, at least without legislative

---

32. “Fair” from both a commercial perspective and “fair” as the term is understood in the context of social justice and equity. *See, e.g.,* Benjamin L. W. Sobel, *Artificial Intelligence’s Fair Use Crisis*, 41 COLUM. J.L. & ARTS 1, 47–48 (2017–2018) (arguing that “unauthorized use of copyrighted data for the sole purpose of debiasing an expressive program” should fall under fair use protections); Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence’s Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018) (exploring how copyright law could improve and exacerbate bias within machine learning systems).

33. *See infra* Part II.A (discussing this doctrinal distinction).

intervention.<sup>34</sup> While we share some of the concerns about the uses to which ML systems may be put, copyright is not the right tool to regulate those abuses.

The implications of this debate go beyond machine learning. Understanding why the use of copyrighted works by ML systems should be fair actually reveals a significant issue at the heart of all copyright law. People, like machines, often copy expression when they are only interested in learning the ideas conveyed by that expression. That's true of most ML training protocols. The ML system wants photos of stop signs so it can learn to recognize stop signs, not because of the artistic choices you made in lighting or composing your photo.<sup>35</sup> Similarly, a natural language generation system wants to see what you wrote to learn how words are sequenced in ordinary conversation, not because it finds your prose particularly expressive or because it wants to use your turn of phrase.<sup>36</sup>

---

34. Cf. *Keller v. Elec. Arts, Inc.*, \_\_\_ F.3d \_\_\_ (9<sup>th</sup> Cir. 2013) (holding that thousands of former college athletes each had the right to prevent EA from making a college football video game; EA shut down the game altogether rather than try to get permission from all the former players).

35. There is a sense in which the creative choices matter even here. The AI is likely to want to see photos of stop signs in a variety of lights, angles, and conditions to train better. But it doesn't value the creativity as creativity.

36. Some systems blur this line. Natural language generators do want to know how words are sequenced together in ordinary human speech, so they may be interested in the way a particular text phrases things. But it is only short phrases that are likely to be relevant to the ML system, and short

ML systems are not alone in wanting works for reasons that have little to do with the authors' actual expression. The issue arises in lots of other copying contexts. In *American Geophysical Union v. Texaco*, for example, the defendants were interested only in the ideas in scientific journal articles; photocopying the article was simply the most convenient way of gaining access to those ideas.<sup>37</sup> Other examples include two pending Supreme Court cases, copyright disputes over software interoperability cases like *Lotus v. Borland*<sup>38</sup> and *Google v. Oracle*,<sup>39</sup> current disputes over copyright in state statutes and rules adopted into law,<sup>40</sup> and perhaps even the tangled morass of cases around copyright protection for the artistic aspects of utilitarian works like clothing, bike racks, and even Bikram yoga.<sup>41</sup> In all these cases, copyright law is being used to target defendants who actually want access to

---

phrases are supposed to be uncopyrightable. See 37 C.F.R. § 202.1 Material Not Subject to Copyright (“(a) Words and short phrases such as names, titles, and slogans . . .”). *But see* Hall v. Swift,—F.3d – (9<sup>th</sup> Cir. 2019) (permitting suit against Taylor Swift based on six-word phrase “players gonna play, haters gonna hate”).

37. *Am. Geophysical Union v. Texaco, Inc.*, 60 F.3d 913, 930 (2d Cir. 1994).

38. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 516 U.S. 233 (1996).

39. *Oracle Am., Inc. v. Google Inc* 886 F.3d 1179 (Fed. Cir. 2018), *cert. granted sub nom* Google LLC v. Oracle America, Inc., No.18-956.

40. *Georgia v. Public.Resource.Org, Inc.*, *cert. granted*, 139 S. Ct. 2746 (2019).

41. *Star Athletica, L.L.C. v. Varsity Brands, Inc.*, 137 S. Ct. 1002, 197 L. Ed. 2d 354 (2017); *Bikrams Yoga College of India LP et al v. Evolution Yoga LLC et al*, \_\_ F.3d \_\_ (9<sup>th</sup> Cir. 2015).

something the law is not supposed to protect—the underlying ideas, facts, or functions of the work.

Copyright law should permit copying of works for non-expressive purposes—at least in most circumstances. While some courts have considered these issues under merger or protectability principles, occasionally denying protection altogether to functional works, the real issue in the cases we consider here is usually not that the work as a whole is unprotectable. Rather, the issue is the fit between what the law protects and what the defendant wants. When the defendant copies a work for reasons other than to have access to the protectable expression in that work, fair use should consider under both factors one and two whether the purpose of the defendant’s copying was to appropriate the plaintiff’s expression or just the ideas.

That said, the protection afforded by fair learning should not be unlimited. When learning is done to copy expression, for example, by training an ML system to make a song in the style of Ariana Grande, the question of fair use can—and should—become much tougher. But, as importantly, we don’t want to allow the copyright on the creative pieces to end up controlling the unprotectable elements.

In Part I, we discuss how machines learn and how this new technological paradigm of “machine learning” differs from other more doctrinally familiar paradigms. In Part II, we explain why ML systems might run afoul of

copyright when it comes to compiling training sets and discuss how copyright law today treats such uses (and how it should in the future). Finally, in Part III we discuss the broader implications of our principle of fair learning for a variety of other copyright disputes.

### I. The Voracious Learner

The last decade has seen the subfield of computer science known as machine learning (ML) take the world by storm.<sup>42</sup> This Part briefly overviews the technological constituents of ML, distinguishing it from “text data mining” (TDM) technologies that copyright law has dealt with—with varying degrees of success—in the last several decades. It then outlines some of the implications ML technologies pose for copyright law,<sup>43</sup> laying the

---

42. For a short overview of this history, see, e.g. Casey & Lemley, *You Might be a Robot*, *supra* note \_\_, at (manuscript at 11–13) (tracing the rise of ML after AI’s so-called “winter”).

43. Numerous previous works that have focused on the subject of AI authorship. *See e.g.*, SIXTY-EIGHTH ANNUAL REPORT OF THE REGISTER OF COPYRIGHTS FOR THE FISCAL YEAR ENDING (Jun 30, 1965) at 4 (offering cursory discussion of computer-generated works); Pamela Samuelson, *The Future of Software Protection: Allocating Ownership Rights in Computer-Generated Works*, 47 U. PITT. L. REV. 1185 (1986); William Ralston, *Copyright in Computer-Composed Music: Hal Meets Handel*, 52 J. COPYRIGHT SOC’Y USA 281 (2005); James Grimmelmann, *There’s No Such Thing as a Computer-Authored Work - and It’s a Good Thing*; Bruce E. Boyden, *Emergent Works*, 39 COLUM. J.L. & ARTS 377, 378 (2015); Annemarie Bridy, *Coding Creativity: Copyright and the Artificially Intelligent Author*, 5 STAN. TECH. L. REV. (2012). Comparatively little attention, however, has been paid to copyright protections afforded to the

foundation for a more substantive discussion of ML and copyright's "fair use"<sup>44</sup> doctrine in the parts that follow.

A. *From Readers to Learners*

It's no exaggeration to describe the last three decades as dominated by robot, not human, readership. Beginning with the advent of internet technologies in the 1980s and 1990s,<sup>45</sup> the world saw its information go from mostly physical to mostly digital.<sup>46</sup> This trend had a certain inexorable logic to it. Instead of locking humanity's collective knowledge on library shelves or magnetic film rolls, digitization allowed us to interact with information without being throttled by the processing speeds of the physical world. Once

---

collections of training data necessary to power these systems. *But see* Sag, *supra* note 48, at 3 (briefly discussing ML technologies but focusing primarily on expert-based TDM systems); Grimmelmann, *Copyright for Literate Robots*, *supra* note \_\_ (focusing primarily on expert-based TDM systems).

44. See Part II *infra* for a fully-fledged definition of "fair use."

45. See Grimmelmann, *supra* note 18, at 659 (writing:

In a world of books and other pre-digital technologies, "copyright . . . left reading, listening, and viewing unconstrained." Ordinary acts of reading did not result in any new copies, and hence did not trigger any of the copyright owner's exclusive rights; nor did readers have access to technologies that would have made copying easy.

46. TDM techniques were in use before the advent of the internet. See, e.g., Don R. Swanson, *Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge*, 30 PERSP. BIOLOGY & MED. 7 (1986).



202x]

Fair Learning clean 92620

117

digitized, information could be replicated, transmitted, altered, searched, and analyzed at the click of a button. And that, in turn, gave rise of a powerful class of technologies specializing in doing exactly that,<sup>47</sup> known in some circles by the umbrella term “text data mining” (TDM) tools.<sup>48</sup> Want, for example, to know when the English-speaking world began using the term “robot” in place of the word “automaton” to describe humanoid machines? Doing so requires, first, mining as many written works produced during the relevant time period as possible and then constructing search queries to isolate and identify that data. And that process begins by instructing a robot (or perhaps automaton?) to read all of them.

---

47. See, e.g., Jerome H. Reichman & Ruth L. Okediji, *When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on A Global Scale*, 96 MINN. L. REV. 1362, 1366 (2012) (noting that “The combination of massive storage capacity, powerful data manipulation techniques, and graphical capabilities has revolutionized both how basic research is conducted and how the resulting knowledge is preserved and disseminated in nearly all fields of science”).

48. According to Matthew Sag, “TDM is an umbrella term referring to computational processes for applying structure to unstructured electronic texts and employing statistical methods to discover new information and reveal patterns in the processed data.” Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 6 J. OF THE COPYRIGHT SOC’Y OF THE USA (2019) (manuscript at 4) (citing Eleanor Dickson, et al., *Synthesis of Cross-stakeholder Perspectives on Text Data Mining with Use-limited Data: Setting the Stage for an IMLS National Forum* (2018) at 5) (quotations removed). Sag further clarifies: “The term ‘text’ in this context is broad enough to include fixed images, sound recordings, and audio-visual works.” See Sag, *supra*, at 4.

Each day, with most of us scarcely noticing,<sup>49</sup> TDM technologies of this variety read and index untold terabytes of data. Search engines read—or in software parlance “crawl”—all of the internet’s text, images, audio, and video data in hopes of organizing it; social media platforms read and categorize all of our cat photos; the list goes on. So, while the notion of a typical “reader” might invoke an image of a student flipping through the pages of an overgrown textbook, the reality is that modern reading is an overwhelmingly robotic affair. As James Grimmelman observed, “[I]f you count by the total number of words<sup>50</sup> read, robotic reading is now overwhelmingly more common than human.”

Today, some thirty years after the rise of the robotic reader, all signs suggest another paradigm shift in readership is underway—this time with a new class of robotics technologies that is less focused on passively reading information than on actively learning from it. The technology comes from a subfield of computer science, known as machine learning (ML), that exploded onto the scene in the last decade.<sup>51</sup> We have elsewhere provided an

---

49. See Grimmelman, *supra* note 18 at 658 (noting that this process occurs almost invisibly).

50. “Words” is perhaps too narrow a term here and also includes content such as video, audio, and images.

51. TOM M. MITCHELL, MACHINE LEARNING, 1–2 (1997).

overview of ML that we need not revisit in detail here.<sup>52</sup> But to understand precisely why this new type of “robotic learner” is in ascent, it’s worth briefly describing how it differs from the mere “robotic readers”<sup>53</sup> that came before it.

Though sometimes conflated with the TDM tools of yesteryear,<sup>54</sup> ML technologies differ in at least one fundamental regard. Unlike “expert”<sup>55</sup> TDM systems that actually slavishly<sup>56</sup> follow a set of rules engineers have hard coded for them, ML systems approach problem-solving tasks in much the same way humans do: by trying to learn. Rather than attempting to imagine and encode every step involved in distinguishing a cat photo from a dog photo, engineers instead create an environment where robots can develop

---

52. See, e.g., Casey & Lemley, *supra* note 7 (overviewing the technology’s fundamentals); Casey & Lemley, *supra* note 7 (same).

53. See generally Grimmelmann, *supra* note 18 (referring to expert-based systems by this coinage).

54. While the Venn Diagrams of TDM tools and ML systems do overlap if one relies on a broad enough definition of TDM, we’re referring to TDM tools that rely on explicit rules-based approaches to mining their insights.

55. The term “expert,” here, refers to a software system programmed to follow an explicit set of hard coded rules without undergoing a process of learning an implicit set of rules through learning.

56. L. Batlin & Son, Inc. v. Snyder, 536 F.2d 486, 490 (2d Cir. 1976) (quoting MELVILLE B. NIMMER, NIMMER ON COPYRIGHT § 6, at 10.2 (1975)).

their own rules for making the distinction through trial and error. An ML system trying to recognize cats, for instance, would be exposed to a series of examples. And with each attempt, the system would improve its chances on the next one by learning from its successes and failures.

Freed from the limitations of an expert-based approach, ML systems have proven capable of reaching heights of achievement far beyond TDM. Whereas TDM systems might effectively catalogue the world's radiology images and perhaps make them searchable by humans, ML systems can now outperform human radiologists at diagnosing medical ailments from them. And it's this ability to learn instead of simply following explicit instructions that has proven central to unlocking vast machine potential.

Yet, while ML approaches resemble human learning in many ways, there's at least one aspect in which humans and machines radically diverge: Machines, it turns out, are remarkably slow learners. To make the judgments we want them to make, ML systems must first be exposed to thousands, millions, or even billions of examples—all of which are collected and stored in a database known as a “training set.” And unlike some TDM systems, which may make only transitory copies,<sup>57</sup> ML systems generally require a more permanent training data set to test successive iterations of the software

---

57. Carroll, *supra* note 106. As Carroll discusses, the legal status of temporary copies is uncertain under copyright law. *Id.*

against. With a large enough training set,<sup>58</sup> virtually no problem-solving task is insurmountable. As we saw in the introduction, machines can compose music,<sup>59</sup> generate artwork,<sup>60</sup> play complex strategy games,<sup>61</sup> and outperform human professionals at the pinnacles of their careers.<sup>62</sup> And all signs suggest this is just the beginning of the technology's long arc of progress.

---

58. And enough computing resources, of course.

59. See *supra* notes 2–5 and accompanying text. See also, e.g., Lori Dorn, *You Can't Take My Door, A Country Song Created by a Neural Network That Studied a Catalog of Country Hits*, LAUGHING SQUID (Apr. 15, 2019), <https://laughingsquid.com/country-song-created-by-neural-network/>.

60. See Mike Murphy, *Computers Can Now Paint like Van Gogh and Picasso*, Quartz (Sept. 6, 2015), <https://qz.com/495614/computers-can-now-paint-like-van-gogh-and-picasso/>. See also, e.g., Gabe Cohn, *AI Art at Christie's Sells for \$432,500*, N.Y. Times (Oct. 25, 2018), <https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html>.

61. Cade Metz, *In a Huge Breakthrough, Google's AI Beats a Top Player at the Game of Go*, WIRED (Jan. 27, 2016), <https://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/>. “Go” is an ancient Eastern strategy game that is comparable to chess, though far more computationally complex. *Id.* Tom Simonite, *Can Bots Outwit Humans In One Of The Biggest Esports Games?*, WIRED (Jun. 25, 2018), <https://www.wired.com/story/can-bots-outwit-humans-in-one-of-the-biggest-esports-games/>. DOTA is one of the internet's most popular real time strategy games and is more difficult for AI systems than Go or chess.

62. *FDA Approves AI-Powered Diagnostic That Doesn't Need a Doctor's Help*, MIT TECH. REV. (Apr. 11, 2018), <https://www.technologyreview.com/the-download/610853/fda-approves-first-ai-powered-diagnostic-that-doesnt-need-a-doctors-help/>.

There is at least one obstacle standing in the way of ML’s seemingly inexorable learning curve. Virtually all the data used to compile training sets is protected by copyright. And just as was true of TDM readers in the 90s, 00s, and 10s, this new breed of robotic readers appears destined to give rise to a host of doctrinal and policy challenges in the years ahead. (Indeed, it already has.)

*B. Copyrights, Copyrights Everywhere*

It’s shockingly easy to create a copyrighted work. Copyrights cover a broad swath of creations, from the written word to art of all types to software, dance, and even architecture.<sup>63</sup> The standard for establishing copyright protection is low—you need only an “original work of authorship” and record it in some more-than-transitory form.<sup>64</sup> And both of those requirements are so trivial as to be almost meaningless. Anything you write longer than a sentence—and art as simple as a red canvas or an accidental stroke of the pen—is likely to meet the originality standard.<sup>65</sup> Moreover, a work can be

---

63. 17 U.S.C. §102(a). The outer limits of copyright law involve things like gardens that evolve on their own. *Kelley v. Chi. Park Dist.*, 635 F.3d 290 (7th Cir. 2011) (holding that a park was not “fixed” and therefore ineligible for copyright protection).

64. 17 U.S.C. §102(a).

65. *See* *Bleistein v. Donaldson Lithographic Co.*, 188 U.S. 239 (1903); *Alfred Bell v. Catalda Fine Arts*, 191 F.2d 99 (2d Cir. 1951).

“fixed” in almost any form, including a simultaneous recording.<sup>66</sup> Once you’ve created the work, no special formalities or applications need be dispatched to protect it. Copyright protections apply immediately.<sup>67</sup>

The threshold is so low, in fact, that it is virtually impossible to go a day without creating multiple copyrighted works. You’d have to stay off email and social media entirely, abstain from selfies or videos, write only extremely short texts, and refrain from doodling. Even that might not be enough. If you have Alexa turned on in your house, your conversations may well be copyrightable.

Once you’ve got all these copyrights, they turn out to be surprisingly hard to undo. Copyright lasts for the life of the author plus 70 years,<sup>68</sup> so even things created in the mid-1920s are still subject to copyright protection. There is no clear means for abandoning a copyright or dedicating it to the public

---

66. See *MAI Sys. Corp. v. Peak Computer, Inc.*, 991 F.2d 511, 513 (9th Cir. 1993) (establishing this broad standard for fixed forms); *NFLC, Inc. v. Devcom Mid-Am.*, 45 F.3d 231, 235 (7th Cir. 1995); *Stenograph v. Bossard Assocs., Inc.*, 144 F.3d 96, 102 (D.C. Cir. 1998). *Contra* *CoStar Grp., Inc., v. LoopNet, Inc.*, 373 F.3d 544, 546-47 (4th Cir. 2004); *Cartoon Network LP v. CSC Holdings, Inc.*, 536 F.3d 121, 129-30 (2d Cir. 2008).

67. See 17 U.S.C. §302(a) (establishing that “[c]opyright in a work . . . subsists from its creation”). Cf. Chris Sprigman, *Reform(aliz)ing Copyright*, 57 *STAN. L. REV.* 485 (2004) (arguing that we should bring back formalities to avoid accidental infringement).

68. See 17 U.S.C. §302(a) (“copyright endures for a term consisting of the life of the last surviving author and 70 years after such last surviving author’s death”).

domain.<sup>69</sup> And even if you do manage to license it to the public at large, perhaps via an open source or creative commons license designed to get around this difficulty, you can always change your mind 35 years down the line and get your copyright back, and no agreement to the contrary can stop you.<sup>70</sup>

The combination of all these factors—a broad range of things protected, the very low standard for copyrightability, the long life of copyrights, and the inability to disclaim them—means that literally tens of billions of new copyrighted works are created every day, and an almost uncountable number of things are copyrighted. Most of them, of course, are worthless. But they’re still protected by federal law.

---

69. See, e.g., Jerry Brito & Bridget Dooling, *An Orphan Works Affirmative Defense to Copyright Infringement Actions*, 12 Mich. TELECOMM. & TECH. L. REV. 75, 76 (2005) (describing how copyright law “renders untouchable a large swath of existing artistic, literary, and other works because if a work’s copyright owner cannot be found to secure their permission to use the work, then no one will ultimately use the work lest they risk liability for copyright infringement”); Olive Huang, *U.S. Copyright Office Orphan Works Inquiry: Finding Homes for the Orphans*, 21 BERKELEY TECH. L. J. 265 (2006) (same); Dave Fagundes & Aaron Perzanowski, *Abandoning Copyrights* (working paper 2020).

70. See generally Peter S. Menell & David Nimmer, *Pooh-Poohing Copyright Law’s “Inalienable” Termination Rights*, 57 J. COPYRIGHT SOC’Y U.S.A. 799 (2009-2010) (discussing the implication of the termination of transfer right).



202x]

Fair Learning clean 92620

125

Given this broad definition, virtually all the training sets used by ML systems include copyrighted works. Object recognition tools and optical scanning software need to train on photographs. So do self-driving cars, self-flying planes, warehouse robots, and any other entity that needs to identify and navigate around obstacles. Speech-recognition and speech-generation systems need to train on recorded audio inputs from radio, TV, movies, and everyday conversations, and virtually all of those were created in the last hundred years and are potentially subject to copyright. Text generation and translation software similarly need to train on a corpus of written works, and if you don't want your texts to be filled with "thous" and "dosts" that means training on works written within the last hundred years. All those things are copyrightable.

Fortunately, there are a number of doctrinal protections afforded to robot readers. The first, and most superficial, is that facts themselves are not subject to copyright protection.<sup>71</sup> So an ML system that needs only pure facts (say, a stock trading algorithm that only studies prior stock purchases) might seem to be off the hook. But a compilation of uncopyrightable facts can itself be copyrighted as long as there is even minimal originality in the selection or

---

71. *See, e.g.*, 17 U.S.C. § 102(b); *Feist Publications, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340 ("copyright protection extends only to those components of the work that are original to the author, not to the facts themselves").

arrangement of those facts.<sup>72</sup> Thus, many databases can be copyrighted even though the individual pieces of data within them are not protected.<sup>73</sup> A database that is comprehensive—including everything in the field—and not creatively organized won't get protection,<sup>74</sup> but many databases will. This includes things we don't normally think of as databases, like medical admissions forms or police booking sheets, each of which selects data about its subjects in an arguably creative way.<sup>75</sup>

---

72. *Id.*

73. *See, e.g.,* CCC Info. Servs., Inc. v. Maclean Hunter Mkt. Reports, Inc., 44 F.3d 61, 63 (2d Cir. 1994) (reversing district court decision finding list of estimated used car valuations to be unprotectable facts and holding that the Red Book numbers as well as the selection and arrangement of the Red Book to be protectable expression); CDN Inc. v. Kapes, 197 F.3d 1256 (9th Cir. 1999) (holding that a list of wholesale prices for rare coins listed by publisher contained sufficient originality to attract copyright protection).

74. Though many of those may effectively get copyright-like protection through the enforcement of ubiquitous terms of use. *See generally* Mark A. Lemley, *Terms of Use*, 91 MINN. L. REV. 459 (2006) (describing this phenomenon). The case that started the “no need to agree” revolution in contract law, was itself a case in which the defendant copied a database that contained only uncopyrightable facts and no creative selection or arrangement. ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1448–49 (7th Cir. 1996). Nonetheless, Zeidenberg was held liable for copying the “unprotectable” database. *See id.*

75. *Compare* Practice Mgmt. Info. Corp. v. Am. Med. Ass'n, 121 F.3d 516, 517 (9th Cir. 1997), *amended by* 133 F.3d 1140 (9th Cir. 1998) (holding federal agency's adoption of work as standard in preparation of Medicare and Medicaid claims did not render copyright invalid), *with* Bibbero Sys., Inc. v. Colwell Sys., Inc., 893 F.2d 1104 (9th Cir. 1990) (medical insurance form didn't have

202x]

Fair Learning clean 92620

127

The copyrightability of databases presents a hurdle for ML systems. But this hurdle is not insurmountable. An engineer or company that wants to train its system using a comprehensive existing database will sometimes (but not always) be able to license that database, and it seems reasonable that it should have to. Precisely because the database collects a lot of information, it offers a valuable form of “one-stop shopping” for ML systems.

There are circumstances where that will not be true. Companies might not voluntarily license their databases to competitors. That’s a problem only if the company that owns the database has exclusive access to the type, volume, or quality of the data stored in the database. That will be true sometimes—think Google’s database of stop signs and its self-driving-car business—but not always. And it should worry us only if we think there are policy reasons to make sure that each competitor has the best possible training data. There’s a good argument for that in some sectors that involve public health and safety, such as medical technologies and safety-promoting technologies (including, perhaps, self-driving cars).

There are also policy concerns related to the transparency and accuracy of the algorithms that employ them. We might want to know what’s in our

---

sufficient creativity because the categories selected were dictated by Medicare and Medicaid); *Southco, Inc. v. Kanebridge Corp.*, 390 F.3d 276 (3d Cir. 2004) (en banc) (holding part numbers used to identify and distinguish among types of screw fasteners not protectable).

training sets where their use has public policy implications. Database transparency, for example, might allow us to figure out whether a photo recognition algorithm is bad at identifying minorities,<sup>76</sup> a criminal sentencing algorithm replicates racial bias,<sup>77</sup> a credit rating algorithm uses potentially biased data,<sup>78</sup> or the like. In those cases, we might want to open access to the database to regulators or compel licensing of the databases to competitors on reasonable terms, just as copyright does in myriad other areas.<sup>79</sup>

But copyright in databases is only a small part of the problem ML systems face. A much more significant and less tractable problem is the

---

76. See, e.g., Jacob Snow, *Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots*, ACLU BLOG (Jul. 26, 2018), <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28> (highlighting this potential, albeit while using a dubious methodology).

77. See, e.g., Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (highlighting this potential, albeit while using a dubious methodology).

78. See, e.g., Kaveh Waddell, *How Algorithms Can Bring Down Minorities' Credit Scores*, THE ATLANTIC (Dec. 2, 2016), <https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/> (describing the potential of this phenomenon in machine learning).

79. See, e.g., 17 U.S.C. §§ 111, 114–115, 119. While property rights advocates worry that these compulsory licensing schemes prevent bargaining and therefore undermine the incentives to create, the evidence doesn't support that in copyright. See Mark A. Lemley, *Contracting around Liability Rules*, 100 CALIF. L. REV. 463 (2012).

202x]

Fair Learning clean 92620

129

copyright in the individual components of the database. It is one thing to license a database of photos of faces from the entity that compiled it. It is quite another to try to get the rights to each individual photo from the millions (or hundreds of millions) of individuals who took them in the first place. ML engineers or companies that want to compile a training set of all books, or all music, or all video content face a similar problem. While books and music tend to have more concentrated copyright ownership, there are still millions of authors and artists and thousands of commercial publishers out there.

Nor will it matter that an ML company didn't get those books or images directly from the copyright owner. Copyright is a strict liability offense.<sup>80</sup> Acting reasonably in getting a license from the database owner won't help you if the database owner doesn't have a license for each and every one of the hundreds of millions of works, even if they claim they do.<sup>81</sup> And even if they do have a license, that license might not cover all of the rights needed

---

80. See R. Anthony Reese, *Innocent Infringement in U.S. Copyright Law: A History*, 30 COLUM. J.L. & ARTS. 133 (2007) (tracing the history of copyright infringement that lacked intent); Patrick R. Goold, *Is Copyright Infringement a Strict Liability Tort?*, 30 BERKELEY TECH. L.J. 305 (2015) (further exploring this phenomenon in the context of tort law).

81. See *Lipton v. Nature Co.*, 71 F.3d 464, 467 (2d Cir. 1995).

from all the owners,<sup>82</sup> or it might be limited to uses that do not include the previously-uncontemplated use by an ML system in training its algorithm.

Nor will it matter that the work is used only inside the ML system and isn't copied in a final output. Intermediate copying is still copying and can still be infringing.<sup>83</sup> It also won't matter that no one book, song, or image has much value to the ML system in the course of its training. If the company deploying the ML application could identify in advance all the owners of all

---

82. To take just one example, streaming a song requires multiple different licenses for rights usually held by different entities: the public performance right in the sound recording, the reproduction right in the sound recording, and the public performance right in the underlying musical composition. See Jason Koransky, *Digital Dilemmas: The Music Industry Confronts Licensing for On-Demand Streaming Services*, AM. BAR ASS'N (JAN. 2016), [https://www.americanbar.org/groups/intellectual\\_property\\_law/publications/landslide/2015-16/january-february/digital-dilemmas-music-industry-confronts-licensing-on-demand-streaming-services/](https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2015-16/january-february/digital-dilemmas-music-industry-confronts-licensing-on-demand-streaming-services/) (describing the complexities of music licensing); Mark A. Lemley, *Dealing with Overlapping Copyrights on the Internet*, 22 U. DAYTON L. REV. 547 (1997).

83. Walker v. University Books, 602 F.2d 859, 864 (9th Cir. 1979) (“[T]he fact that an allegedly infringing copy of a protected work may itself be only an inchoate representation of some final product to be marketed commercially does not in itself negate the possibility of infringement.”); Sega of Am. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992) (“intermediate copying of computer object code may infringe the exclusive rights granted to the copyright owner in section 106 of the Copyright Act regardless of whether the end product of the copying also infringes those rights.”). For a discussion of that principle applied to AI-generated art, concluding that creating intermediate reproductions is an act of infringement by an AI, see Jessica L. Gillotte, *Copyright Infringement in AI-Generated Artworks*, 53 U.C. DAVIS L. REV. 2655, 3672-73 (2020).

the works it will use, perhaps it could negotiate licenses with all of them or exclude the works for which it couldn't obtain a license. But that is impractical. Most photographs, for instance, have no copyright management information attached to them. It is effectively impossible to find that many copyright owners and negotiate that many licenses. That's why Congress created compulsory licenses for many modern uses, and why private groups like ASCAP and BMI organized to collect and license copyrights in specific sectors. Perhaps eventually similar organizations will arise to license individual works for training datasets. But they don't exist now, and the fact that the training set needs to be comprehensive means it will be a long time before anyone could effectively aggregate all the rights needed.

Further, the fact that no one work is valuable in the training process won't prevent copyright lawsuits. Copyright remedies are structured to encourage lawsuits over even small-value infringements. Copyright law awards statutory damages of up to \$150,000 per work, plus attorneys' fees, regardless of the plaintiff's actual loss or the defendant's actual gain.<sup>84</sup> True, statutory damages require registration of the work,<sup>85</sup> and many private citizens won't have done that. But most copyrighted books and songs and many commercial videos and photographs are registered, so the risk is greater

---

84. 17 U.S.C. § 504(c)(2).

85. 17 U.S.C. § 412 (establishing registration as a prerequisite for such damages),

for databases that build on those works. Multiply that by the number of works at issue and the risk to the ML systems of using training datasets featuring potentially infringing works becomes enormous.

Given these incentives, it's not difficult to foresee some plaintiffs suing opportunistically, just to collect the potential windfall. In addition, we're increasingly seeing copyright law being floated as a tool to prevent ML access for noneconomic reasons. The continued expansion of ML systems into virtually all walks of social and commercial life has raised a new set of concerns around the protections that copyright affords such technologies.<sup>86</sup> With ML systems entering our roadways, courtrooms, and police stations, headlines decrying new applications that push data usage norms to the limits have become regular features of the news.<sup>87</sup> With few clear legal mechanisms available to potential plaintiffs, the question of whether copyright might be a useful tool for curbing the proliferation of such abuses has come to center stage. Today, copyright's future regarding ML systems remains far from certain. But what is certain is that, as ML's capabilities grow with time, the temptation to transform copyright into a tool to restrict them will too.

---

86. *See infra* Part II.

87. *See supra* notes 10–17 and accompanying text.



## II. Fair Learning

In Part I, we saw that ML requires access to unprecedented amounts of information. Most, if not all, of that information will be copyrighted. And those copyrights are likely to be owned by thousands or even millions of different parties. At least in theory, that creates a problem for training ML systems. We explain why it hasn't yet served as a roadblock for ML progress in subpart A. In subpart B, we explore some of the new legal challenges that arise when robots attempt to learn from, rather than merely search, copyrighted material. In subpart C, we explore how current fair use doctrine is coming under pressure in the context of ML. And, finally, in subpart D we propose a standard that courts and technologies can use to get out of the hole that recent copyright precedent and increasingly strident public opinion have dug for them.

### *A. Copyright for Robotic Readers*

With all the copyright pitfalls awaiting anyone attempting to create an ML training set, one might wonder why copyright law hasn't effectively halted the gears of progress in the field. The oversimplified answer needs only two words: "fair use."<sup>88</sup> The fair use doctrine permits certain uses of

---

88. 17 U.S.C. §107.

copyrighted works for valuable social purposes,<sup>89</sup> particularly when such uses “transform” the original source material<sup>90</sup> and do not threaten the copyright owner’s core market.<sup>91</sup> In Judge Pierre Leval’s famous articulation:

The use must be productive and must employ the quoted matter in a different manner or for a different purpose from the original[.] If . . . the secondary use adds value to the original – if the quoted matter is used as raw material, *transformed* in the creation of new information, new aesthetics, new insights and understandings – this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society.<sup>92</sup>

Since Judge Leval penned those words more than three decades ago, his description of transformative fair use has, itself, transformed—growing to

---

89. See, e.g., ABRAHAM DRASSINOWER, WHAT’S WRONG WITH COPYING? 78 (2015) (“[T]he defense is not about undoing or overlooking a wrong for reasons extraneous to authorship itself. . . . It is as if, upon hearing the plaintiff’s complaint, the defendant were to say: ‘. . . I am equally an author.’”).

90. Authors Guild v. Google, Inc., 804 F.3d 202, 214-15 (2d. Cir. 2015). See, e.g., Asay *supra* note 8 (describing the trend).

91. 17 U.S.C. §107(4).

92. Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) (emphasis added). Fair use, of course, didn’t originate with Judge Leval. See, e.g., Sag, *supra* note 48, at 10 (noting that “[f]air use and its historical antecedents have been part of copyright law since very shortly after the enactment of the first copyright act, the English Statute of Anne in 1710”).

202x]

Fair Learning clean 92620

135

encompass not just a host of human-authored changes to the copyrighted works (e.g. parodies, critiques, fan fiction, and the like) but robotic ones as well.<sup>93</sup> Among the most significant moments in this trajectory was *Sega Enterprises Ltd. v. Accolade, Inc.*<sup>94</sup> The controversy began when the video game publisher, Accolade, decided it wanted its games to run on Sega's wildly popular Genesis console. To do so, Accolade needed its software to mirror the communication protocols Sega used to interact with the console. And the only way to unlock those protocols was by copying large parts of Sega's software verbatim to reverse engineer it. When Sega caught wind of Accolade's efforts, it brought suit—arguing that the conduct violated its copyright in the video game software.<sup>95</sup> Accolade countered that its direct copying was necessary as an intermediate step toward accessing the unprotectable “ideas and functional elements” hidden in Sega's object code.<sup>96</sup>

---

93. Clark D. Asay & Arielle Sloan, *Is Transformative Use Eating the World?*, 61 B.C. L. REV. (forthcoming 2020).

94. 977 F.2d 1510 (9th Cir. 1992).

95. Because the question of whether such protocols were protected by trade secrets had already been resolved, *see e.g. Acuson Corp. v. Aloka Co., Ltd.*, 257 Cal. Rptr. 368, 10 U.S.P.Q.2d (BNA) 1814 (App. 6th Dist. 1989), the case instead centered on whether copyright's fair use doctrine protected it.

96. *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992).

Given that Accolade had copied significant amounts of Sega's work without commenting on, or modifying, the original expression, the case wasn't exactly clear cut (at least, under Judge Leval's original articulation of transformative use).<sup>97</sup> But the Ninth Circuit found the use fair, drawing a "distinction between the copying of works in order to make independent creative expression possible and the simple exploitation of another's creative efforts."<sup>98</sup> Accolade had "copied Sega's code for a legitimate, essentially non-exploitative purpose" because the act of copying was necessary as an intermediate step to access unprotectable elements of the work.<sup>99</sup> The court concluded that so long as copying served simply as a means of accessing ideas, facts, or functionality—and not the original expression of the work—the use was fair game.

Since *Sega*, the legal system has gone on to clarify and expand the protections afforded to this type of copying, now commonly referred to as "non-expressive"<sup>100</sup> use. Courts have held in favor of video game

---

97. For a more straightforward case under Judge Leval's articulation, see, e.g., *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 583 (1994) (holding a rap version of a popular song was a parody and potentially transformative under fair use analysis).

98. *Sega*, 977 F.2d 1510 at 1523.

99. *See id.* at \_\_.

100. *See, e.g., Sag, supra* note 48, at 6 (describing the use as "non-expressive"); Grimmelmann, *Copyright for Literate Robots, supra* note 18, at \_\_ (same); Sobel, *supra* note 32 at \_\_ (same).

202x]

Fair Learning clean 92620

137

“emulators” that copied console firmware in order to access its underlying functionality,<sup>101</sup> search engine providers that bulk-collected and displayed images in order to make them more readily accessible,<sup>102</sup> and even plagiarism detection providers that consumed copyrighted materials to use and improve their software.<sup>103</sup> Most famously, courts permitted Google to scan all the world’s books into its internal database as an intermediate step toward producing a book search system that, then, delivered verbatim “snippets” of copyright text to the reader.<sup>104</sup> (More on this below.)

Non-expressive use protections like these (along with the Digital Millennium Copyright Act)<sup>105</sup> are the reason most automated search and analysis tools exist in the first place. Without such protections, Google and

---

101. Sony Computer Entertainment, Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).

102. See Kelly v. Arriba Soft Corp., 280 F.3d 934 (9th Cir. 2002) (finding fair use); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007) (finding fair use).

103. A.V. ex rel. Vanderhuy v. iParadigms, LLC, 562 F.3d 630, 645 (4th Cir. 2009) (holding that copying a student paper to search it for plagiarism had an entirely different function and purpose than the original work.).

104. Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir.2014); Authors Guild v. Google, Inc., 804 F. 3d 202 (2d Cir. 2015).

105. 17 U.S.C. §512. That statute allows sites like YouTube to host (and therefore to index and search) video content uploaded by third parties, so long as they take down infringing material when the copyright owner notifies them of infringement. Without that statutory exception video sites like YouTube would not exist.

others wouldn't be able to copy the large bodies of text and images necessary to make them searchable in the first place.<sup>106</sup> But even this favorable line of precedent doesn't mean that Google or any other technology company has free rein with data it has collected. It is still subject to a fact-specific four-factor test.<sup>107</sup> And, like any fact-specific test, it's unpredictable. In fact, it is so uncertain that Larry Lessig famously described fair use as nothing more than the right to hire a lawyer.<sup>108</sup>

The good news is that most companies deploying ML systems can afford to hire lawyers. But while precedent like *Kelly v. Arriba Soft* and *Author's Guild v. Google* may look promising for those who collect and use

---

106. See *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015). For an argument that text data mining should be fair use even if the TDM system doesn't have lawful access to the text, see Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 U.C. DAVIS L. REV. 893 (2019). For an argument that verbatim copying necessary to create a database should be fair use so long as the output of that database uses only a limited amount of the content, see Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 846 (2010).

107. See Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005*, 156 U. PA. L. REV. 549 (2008); David Nimmer, "Fairest of them All" and Other Fairy Tales of Fair Use, 66 L. AND CONTEMPORARY PROBLEMS 263-288 (Winter 2003); David Nimmer, *Juries And The Development Of Fair Use Standards*, 31 HARV. J. L. & TECH. 563 (2018).

108. Lawrence Lessig, *FREE CULTURE* 187 (2004).

202x]

Fair Learning clean 92620

139

datasets,<sup>109</sup> there is no guarantee that courts will extend these precedents to similar technologies or legal contexts. Indeed, as we'll see in the subpart that follows, both the courts and the court of public opinion have begun to depart from the precedents established by the Google Books cases. And these departures could have lasting ramifications for the use of copyrighted data to train ML systems.

*B. Copyright for Robot Learners*

There is reason to worry that courts won't find AI learning to be fair. Courts applying existing law to ML systems might plausibly conclude that several of the fair use factors weigh against fair use of individual works in training datasets.<sup>110</sup> First, much like TDM systems, ML systems involve copying the entire work without alteration. That directly affects statutory factor number three, which weighs the fact that the entire work is taken

---

109. Grimmelman goes so far as to say that courts ignore bulk collection by robots, *see* Grimmelman, *supra* note 18 at 674 (stating, "Copyright ignores robots.") but that's too strong. Rather, courts focus on what is done with the material after it is collected to influence their assessment of the legality of collecting it.

110. Benjamin Sobel comes to the same conclusion. *See* Sobel, *supra* note 32. Our focus here is on the copyright claims by individual plaintiffs rather than a claim involving wholesale copying of a copyrighted database itself. We think the latter is (and should be) much less likely to qualify as fair use.

against a finding of fair use.<sup>111</sup> But this act of direct copying also affects whether the use is transformative. In *Kelly*, for instance, the court focused not on the reduced resolution of thumbnail images but on the fact that thumbnail images couldn't substitute for full-size images (at least on 2002-era devices) and served a very different purpose.<sup>112</sup>

The closest analogies to the type of direct copying involved in the creation of a training set may seem to be the intermediate copying software cases running from *Sega* to *Google Books*.<sup>113</sup> Notably, however, some of

---

111. Indeed, some judges have said that they would declare wholesale copying to be illegal in all circumstances. *Texaco, Inc.*, 60 F.3d at 930. Fortunately, that is not the law.

<sup>112</sup> *Kelly v. Arriba Soft Corp.*, 280 F.3d 934 (9th Cir. 2002) (finding fair use); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146 (9th Cir. 2007) (finding fair use). Gilotte argues that AI training data sets are “highly transformative” because the purpose of the use is different – to train an AI. Gilotte, *supra* note \_\_, at 2684. We discuss that different purpose in more detail *infra* notes \_\_-\_\_ and accompanying text.

113. As discussed in a series of cases, the Second and Ninth Circuits have held that it is fair use to copy the entire body of a work that includes copyrightable expression as an intermediate step towards creating an end product that does not include that copyrightable expression. The Ninth Circuit cases involve reverse engineering computer programs to produce a different program that is interoperable with the original. That process involves creating an internal working version of the plaintiff's code and then writing new code that correctly interfaces with it. 113. And in *Authors Guild v. Google Inc.*, the Second Circuit held that Google could scan the text of books in order to produce a search engine that could find text in those books and display short snippets of text in response.



202x]

Fair Learning clean 92620

141

these cases have depended heavily on the fact that the defendant's end product was a transformative new work and the copying was a necessary step to get there. Muddying the fair use question further is the fact that several cases—including *Associated Press v. Meltwater U.S. Holdings*<sup>114</sup> and *Fox News v. TVEyes*<sup>115</sup>—have rejected fair use arguments in somewhat analogous contexts.<sup>116</sup> *TVEyes* in particular rejected the district court's finding that a TV news clipping service that analyzed and made TV news searchable was fair use.<sup>117</sup> And the Supreme Court is currently considering its first case involving transformative use since it adopted the doctrine in 1994, so the scope and even the continued existence of the doctrine are up in the air.<sup>118</sup>

---

114. *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 543–44 (S.D.N.Y. 2013).

115. *Fox News Network, LLC v. TVEyes, Inc.*, 883 F. 3d 169 (2nd Cir. 2018).

116. *Meltwater*, 931 F. Supp. 2d at 537; *TVEyes*, 883 F.3d at 169. *But see* Sag, *supra* note 48 at 23–27 (arguing that the *TVEyes* decision does not necessarily suggest that “the tide will turn against TDM and similar non-expressive uses in American copyright law”). We think those cases can be distinguished on their facts—the output of those systems involved sizeable chunks of the original copyrighted work. But it's not clear that they *will* be distinguished.

117. *TVEyes*, 883 F.3d at 169.

118. *Google Inc. v. Oracle of Am.*, 18-956, \_\_ S.Ct. \_\_ (Nov. 15, 2019).

*TVEyes* might be distinguished on its facts,<sup>119</sup> and it's possible that precedents like the Google Books cases will still be extended to ML training in straightforward fashion. But enterprising plaintiff's lawyers can point to a number of facts that might distinguish those cases. The fact that in most cases the ML systems aren't producing a new copyrightable work at all, but are just consuming the plaintiff's work for profit to improve uncopyrighted systems like self-driving cars or speech recognition technology, might persuade a court that those precedents aren't all that helpful.<sup>120</sup> Maybe that won't matter. Like search engines, ML systems use the plaintiff's work only inside the computer, and it is not shared with the public. But it is also possible that the fact that the work is sometimes not communicated to the outside world in any form may hurt rather than help the claim to transformation. Copyright owners may argue that, unlike the search engine cases, some ML systems lack a creative "end product" provided to consumers that is different in nature or purpose than the original work. Rather, the argument goes, ML is a consumer

---

119. Part of what is driving that case is the fact that the defendants ingested the entirety of Fox News in order to create a searchable database of things said there. An AI wouldn't necessarily have to do that in order to learn. But it might well want to if it is to make better decisions and avoid discrimination.

120. *But see* Grimmelmann, *supra* note 18 (arguing, before *TVEyes*, that courts ignore automatic copying by robots). Grimmelmann's focus was on older TDM systems, not the modern ML systems.

202x]

Fair Learning clean 92620

143

of the copyrighted work, outputting a profitable technology rather than new creativity, and these for-profit consumers should pay for what they take.

That argument may appeal to the very strong, if often unarticulated, anti-free riding instinct in courts.<sup>121</sup> And the appeal of that argument is likely to be strengthened by the commercial nature of many ML applications. While commercial uses are not presumptively unfair,<sup>122</sup> they still tend to weigh against a finding of fair use.<sup>123</sup> And commerciality often goes hand in hand with a market effect. Here, the system's use doesn't cut into the ordinary market for the copyrighted works in question. But ML companies might be natural candidates for a licensing market: large for-profit companies that stand to benefit financially from using copyrighted works (albeit in bulk, rather than this work in particular).<sup>124</sup> Given the well-known circularity of the

---

121. See Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEX. L. REV. 1031 (2005); Mark A. Lemley & Mark P. McKenna, *Unfair Disruption*, 100 B.U. L. REV. 71 (2020); [Wendy J. Gordon, \*The Core of Copyright: Authors, Not Publishers\*](#), 52 HOUS. L. REV. 613 (2014).

122. See *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994) (reversing precedent to that effect); *Authors Guild v. Google, Inc.*, 804 F.3d 202, 219 (2d Cir. 2015). (ruling there is “no reason ... why [a defendant's] overall profit motivation should prevail as a reason for denying fair use over its highly convincing transformative purpose, together with the absence of significant substitutive competition, as reasons for granting fair use”).

123. See, e.g., *Sony of Am. v. Universal City Studios*, 464 U.S. 417 (1984).

124. Sobel, *supra* note 32 (discussing this possibility).

claim to a loss of a licensing opportunity—the use is unfair if there is a lost licensing opportunity, but there is only a lost licensing opportunity if the use is unfair<sup>125</sup>—courts may well let their view of the equities creep into the analysis of the fourth factor.<sup>126</sup>

---

125. See Mark A. Lemley, *Should a Licensing Market Require Licensing?*, 70 L. & CONTEMP. PROBS. 185 (2007) (noting this problem); Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1124 (1990) (“By definition, every fair use involves some loss of royalty revenue because the secondary user has not paid royalties.”); William W. Fisher III, *Reconstructing the Fair Use Doctrine*, 101 HARV. L. REV. 1659, 1671 (1988); Sara K. Stadler, *Forging a Truly Utilitarian Copyright*, 91 IOWA L. REV. 609 (2006); Christina Bohannon, *Reclaiming Copyright*, 23 CARDOZO ARTS & ENT. L.J. 567, 597–98 (2006); Matthew Sag, *God in the Machine: A New Structural Analysis of Copyright’s Fair Use Doctrine*, 11 MICH. TELECOMM. & TECH. L. REV. 381, 393–94 (2005); Lydia Pallas Loren, *Redefining the Market Failure Approach to Fair Use in an Era of Copyright Permission Systems*, 5 J. INTELL. PROP. L. 1, 38–39 (1997).

126. Wendy Gordon famously wrote that fair use worked as a substitute where transactions costs made a market infeasible. Wendy J. Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors*, 82 COLUM. L. REV. 1600 (1982). While she has later criticized unwarranted extensions of this work to create spurious markets, Wendy J. Gordon, *Excuse and Justification in the Law of Fair Use: Transaction Costs Have Always Been Part of the Story*, 50 J. COPYRIGHT SOC’Y U.S.A. 149 (2002-2003), plaintiffs have seized on the transactions costs story to create markets for licensing uses that would otherwise be fair. See Lemley, *Licensing Market*, *supra* note (noting this problem).

Jane Ginsburg has argued that the statute makes a use unfair if it cuts into the “value” of the work even if it doesn’t interfere with an established market. Jane C. Ginsburg, *Fair Use Factor Four Revisited: Valuing the “Value of the Copyrighted Work”*, J. COPYRIGHT OFC. SOC’Y (forthcoming

A second challenge for those seeking fair use protections arises in situations where ML systems do, in fact, produce tangible outputs that could trigger a similar strain of anti-free riding sentiment. Particularly in the last several years, we've seen major strides involving ML systems capable of replicating the outputs of creative professionals. MuseNet, discussed in the introduction, is one such example. But ML applications have shown similarly impressive results in fields as diverse as journalism,<sup>127</sup> poetry,<sup>128</sup> painting,<sup>129</sup> and photography.<sup>130</sup> These ML systems—virtually all of which are trained on copyrighted works—have produced writing that's difficult to distinguishing

---

2020). But that too could be circular. A work would be more valuable if ML systems would pay for its use.

127. See *supra* notes \_\_\_ – \_\_\_ and accompanying text.

128. See, e.g., Kelsey Piper, *A Poetry-writing AI Has Just Been Unveiled. It's . . . Pretty Good*, VOX (May 15, 2019), <https://www.vox.com/2019/5/15/18623134/openai-language-ai-gpt2-poetry-try-it>.

129. See Mike Murphy, *Computers Can Now Paint like Van Gogh and Picasso*, Quartz (Sept. 6, 2015), <https://qz.com/495614/computers-can-now-paint-like-van-gogh-and-picasso/>. See also, e.g., Gabe Cohn, *AI Art at Christie's Sells for \$432,500*, N.Y. Times (Oct. 25, 2018), <https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html>.

130. See Samantha Cole, *This Company Promises to Place Any Face Onto Any Body, Using an Algorithm*, Vice (Nov. 21 2019), [https://www.vice.com/en\\_us/article/7x5nv4/rosebud-ai-stock-images-using-an-algorithm](https://www.vice.com/en_us/article/7x5nv4/rosebud-ai-stock-images-using-an-algorithm) (describing a generative system that produces high quality stock photos that are on par with a professional photographer).

from real journalists,<sup>131</sup> painted in the style of celebrated masters,<sup>132</sup> and even created stock photos comparable to those of professional photographers.<sup>133</sup> Many of these efforts have been so convincing that professionals and opinion columnists alike have begun to openly worry about artificial intelligence as a genuine competitive threat. These concerns, in turn, have triggered criticisms from thought leaders, advocates, academics, and professionals worried that technology companies producing these technologies may be free riding on the labor of creative professionals. Critics of such practices have compared leading ML companies to “robber barons” siphoning up valuable IP.<sup>134</sup> Others have vocalized concerns that ML “empowers these companies

---

131. See, e.g., Seabrook, *supra* note \_\_ (describing GPT-2’s linguistic prowess); <https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker>.

132. See Mike Murphy, *Computers Can Now Paint like Van Gogh and Picasso*, QUARTZ (Sept. 6, 2015), <https://qz.com/495614/computers-can-now-paint-like-van-gogh-and-picasso/>. See also, e.g., Gabe Cohn, *AI Art at Christie’s Sells for \$432,500*, N.Y. TIMES (Oct. 25, 2018), <https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html>.

133. See Samantha Cole, *This Company Promises to Place Any Face Onto Any Body, Using an Algorithm*, VICE (Nov. 21 2019), [https://www.vice.com/en\\_us/article/7x5nv4/rosebud-ai-stock-images-using-an-algorithm](https://www.vice.com/en_us/article/7x5nv4/rosebud-ai-stock-images-using-an-algorithm) (describing a generative system that produces high quality stock photos that are on par with a professional photographer).

134. See, e.g., Andrew Orłowski, *Jaron Lanier: Big Tech is worse than Big Oil*, THE REGISTER (Apr. 22, 2019), [https://www.theregister.co.uk/2016/04/22/jaron\\_lanier\\_on\\_ip/](https://www.theregister.co.uk/2016/04/22/jaron_lanier_on_ip/); Solon, *supra* note

202x]

Fair Learning clean 92620

147

to extract value from authors' protected expression without authorization, and to use that value for commercial purposes that may someday jeopardize the livelihoods of human creators."<sup>135</sup> It is not at all clear these practices are reflective of the kind of exploitation of original expression that copyright law is meant to guard against.<sup>136</sup> But what *is* clear is that this emerging view of the equities, too, could have consequences for how courts consider the competitive and substitutive implications of a permissive fair use doctrine.<sup>137</sup>

---

26 (quoting one critic describing large-scale data collection as “the money laundering of facial recognition [where companies are] laundering the IP and privacy rights out of the faces.”)

135. *See generally* Sobel, *supra* note 32. Sobel argues that “[c]onstruing fair use to protect this activity will place the doctrine at odds with the public interest and potentially exacerbate the social inequalities that AI threatens. . . [but at] the same time, finding that expressive machine learning is not fair use would frustrate the progress of the promising technology.” He views this inherent tension as a dilemma for fair use doctrine in the future. *See also, e.g., supra* notes 31–32 and accompanying text. In a separate paper, Sobel notes that courts are unlikely to find fair use when an AI creates a new work in the style of an existing artist. Sobel, *Style, infra* note 179, at 58.

136. *See, e.g., Sony Computer Entertainment, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000) (noting there is a “distinction between the copying of works in order to make independent creative expression possible and the simple exploitation of another’s creative efforts”). If a new artist writes a different and better song than you because she learned from you, the law celebrates that creativity. It is not obvious that it should feel differently because it is a machine rather than a new artist that wrote the better song.

137. *See* Part II.D *infra*.

A third challenge facing those advancing fair use arguments will be the host of other potential negative impacts that the technology can have on downstream users or consumers. As ML systems have been handed greater decision-making authority over our social, economic, and political lives, they've also come under increasing fire by critics who fears the prospect of ML systems replacing human decision makers to negative effect. Today, media outlets, thought leaders, and advocacy organizations decrying new ML systems that push data usage norms to the limits have become a regular feature of the news.<sup>138</sup> Critics variously worry that ML systems with free rein

---

138. See, e.g., Rachel Botsman, *Dawn of the Techlash*, GUARDIAN (Feb 10, 2018), <https://www.theguardian.com/commentisfree/2018/feb/11/dawn-of-the-techlash>; Eve Smith, *The Techlash Against Amazon, Facebook and Google—And What They Can Do*, ECONOMIST (Jan. 20, 2018), <https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do>; Gregory Bobillot, *'Techlash' – How Big Tech is Influencing Your Thinking*, FINANCIAL TIMES (Mat 10, 2018), <https://www.ft.com/video/3339f59e-f760-4bc7-b359-3899fabbd190>; *The Techlash Has Just Begun*, AXIOS (Jan. 10, 2018), <https://www.axios.com/the-techlash-1515609266-e27ca299-0031-460a-96f1-db842ec88121.html>; John Rubio, *'Tech Wreck,' 'Techlash,' 'Techmageddon' - Whatever You Call It, Wall Street Is Terrified Of It*, SEEKING ALPHA (Mar. 29, 2018), <https://seekingalpha.com/article/4159827-tech-wreck-techlash-techmageddon-whatever-call-wall-street-terrified>; *Emails Show How Amazon is Selling Facial Recognition System to Law Enforcement*, ACLU OF NOR. CAL. (May 22, 2018), <https://www.aclunc.org/news/emails-show-how-amazon-selling-facial-recognition-system-law-enforcement>.



to consume copyrighted materials could spread propaganda,<sup>139</sup> facilitate dystopian surveillance,<sup>140</sup> invade personal and sexual privacy,<sup>141</sup> perpetuate

---

139. See, e.g., Alex Hern, *New AI Fake Text Generator May Be Too Dangerous to Release, Say Creators*, THE GUARDIAN (Feb. 14, 2019), <https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction>; @BuzzFeed, *You Won't Believe What Obama Says in This Video*, TWITTER (Apr. 17, 2018, 8:00 AM), <https://twitter.com/BuzzFeed/status/986257991799222272> [<https://perma.cc/C38K-B377>] (“We’re entering an era in which our enemies can make anyone say anything at any point in time.”); Tim Mak, *All Things Considered: Technologies to Create Fake Audio and Video Are Quickly Evolving*, NAT’L PUB. RADIO (Apr. 2, 2018), <https://www.npr.org/2018/04/02/598916380/technologies-to-create-fakeaudio-and-video-are-quickly-evolving> (discussing ML systems’ ability to generate videos created for misinformation campaigns).

140. See, e.g., Cade Metz, *Facial Recognition Tech Is Growing Stronger, Thanks to Your Face*, N.Y. TIMES (Jul. 13, 2019), <https://www.nytimes.com/2019/07/13/technology/databases-faces-facial-recognition-technology.html> (citing instances of facial recognition systems tools trained on photos used by Immigration and Customs Enforcement to identify undocumented immigrants and by Chinese government agencies to engage in “ethnic profiling of the country’s minority Uighur Muslims,” among other instances); Josh Kaplan, *License Plate Readers Are Creeping Into Neighborhoods Across the Country*, SLATE (Jul 10, 2019), <https://slate.com/technology/2019/07/automatic-license-plate-readers-hoa-police-openalpr.html> (documenting instances of automated license plate readers being used to as a surveillance tool by law enforcement and landlords alike).

141. David Greene, *We Don’t Need New Laws for Faked Videos, We Already Have Them*, ELECTRONIC FRONTIER FOUNDATION BLOG (Feb. 13, 2018), <https://www EFF.ORG/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them> (floating copyright as a potential tool for curbing the collection of data for ML training sets).

bias,<sup>142</sup> and even threaten to take over the world.<sup>143</sup> As these systems grow to play even greater roles in the most intimate aspects of our lives, these concerns will almost certainly grow too. These concerns don't relate directly to whether a use is fair (though the public interest purpose of a use is one factor the courts consider), but they may incline parties and courts not to give ML systems the benefit of the doubt. That is particularly true when the ML systems are in the hands of established tech giants like Google or Amazon that are already the target of enormous public ire.<sup>144</sup>

Finally, the sheer number of works ML systems must copy means that taking a chance on fair use is a risky move for an ML company. Copyright

---

142. See, e.g., Jacob Snow, *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*, AMERICAN CIVIL LIBERTIES UNION BLOG (Jul. 26, 2018), <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28> (experimenting with facial recognition technology on headshots of Congressmembers and finding that “false matches were disproportionately of people of color”); Natasha Singer & Cade Metz, *Many Facial-Recognition Systems Are Biased, Says US Study*, N.Y. TIMES (Dec. 19, 2019), <https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html> (describing studies showing racial and other bias in facial recognition tools).

143. See, e.g., Grimmelmann, *supra* note 18 (offering a syllogism concluding that a permissive fair use standard “arguably increases the chances that humanity will meet a sudden, violent, and extremely unpleasant end” at the hands of super intelligent machines).

144. For discussion of the backlash against big tech companies, see, e.g., Mark A. Lemley & Andrew McCreary, *Exit Strategy*, \_\_ B.U. L. REV. \_\_ (forthcoming 2021), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3506919](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506919).

202x]

Fair Learning clean 92620

151

statutory damages systematically overcompensate plaintiffs with small-value works, offering them up to \$150,000 per work regardless of the value the user places on using that particular work.<sup>145</sup> An ML that copies millions of works could potentially face hundreds of billions of dollars in statutory damages. And with thousands or even hundreds of thousands of different copyright owners, the risk of multiple opportunistic suits is high. Many ML companies will not share Google's willingness to bet the company on the legal principle of fair use.

We want to be clear: we aren't saying courts, or the court of public opinion, will definitely reject the fair use defense as it is currently understood, only that there is a risk they will do so.<sup>146</sup> And we certainly aren't saying they

---

145. 17 U.S.C. 504.

146. Outside the United States things are much more varied. Consistent with their more high-protectionist view of copyright, especially as applied to technology, European Union courts are much more likely to hold international companies deploying ML systems liable for copyright infringement. *See, e.g.*, *Infopaq International A/S v Danske Dagblades Forening* (2009) (holding that an 11-word snippet displayed in a context similar to Google's search might infringe copyright). In Asia, by contrast, countries are friendlier towards ML. Japan recently adopted a law permitting ML training as an exception to copyright. *See, e.g.*, *Japan Amends Its Copyright Legislation to Meet Future Demands in AI and Big Data*, EURO. ALLIANCE FOR RESEARCH EXCELLENCE (Sep. 3, 2018), <http://eare.eu/japan-amends-tdm-exception-copyright/> (describing Japan's adoption of legislation that explicitly allows copying by ML engineers collecting data for training sets). The European Union permits TDM in some circumstances. Carroll, *supra* note 106, at 895-96. And Singapore is

*should* reject the defense. To the contrary, there are very strong policy reasons to allow ML systems to copy individual works in the course of training algorithms, as we explain in the following section.

C. *Just the Facts, Ma'am*

There are several reasons why the fair use doctrine should permit ML systems to train on data sets that include copyrighted works.

First, society benefits from allowing ML systems to compile the best possible databases and to open them for public scrutiny and for open AI. Broad access to training datasets will make AI better, safer, and fairer.<sup>147</sup> Smaller, proprietary datasets – particularly those with large and non-random gaps due to failures of copyright licensing – will lead to worse decisions by ML systems. And those worse decisions have real-world consequences. They may mean the difference between a self-driving vehicle that stops at a stop sign at night in the rain and one that doesn't. Or the difference between a biometric scanner at airports that accurately identifies women of color and one that doesn't. Or between dictation software that faithfully transcribes

---

considering a similar law. *See, e.g.,* Christian Troncoso, *Copyright Proposal Threatens to Undermine Europe's AI Ambitions*, BSA TECHPOST (Sep. 5, 2018), <https://techpost.bsa.org/2018/09/05/copyright-proposal-threatens-to-undermine-europes-ai-ambitions/> (noting that Singapore courts and legislatures are grappling with the issue of fair use by ML systems).

147. *See supra* note 32 and accompanying text.

202x]

Fair Learning clean 92620

153

what you said and dictation software that lands you in the hall of fame at Autocorrect Fail.<sup>148</sup>

Second, given the large number of works an AI training data set needs to use and the fact that thousands if not millions of different people own those works, AI companies can't simply license all the underlying photographs or texts for the new use. So allowing a copyright claim is tantamount to saying, not that copyright owners will get paid, but that no one will get the benefit of this new use because it will be impractical to make that use at all.<sup>149</sup> That is particularly ironic if the copyright claim is justified by the supposed existence of a licensing market, since the very theory of the licensing market will have blocked the creation of such a market.<sup>150</sup> Or (and this may be just as bad), only companies like Google or Facebook that happened to collect the data for other, permissible purposes will be able to compete in AI space.<sup>151</sup>

---

148. *See, e.g.*, [www.autocorrectfail.com](http://www.autocorrectfail.com). The original, [damnyouautocorrect.com](http://damnyouautocorrect.com), appears to be defunct, alas.

149. *Cf.* Keller, *supra* note 34.

150. For an argument that any plaintiff who relies on a licensing market argument to defeat a fair use claim should have to forego any claim to injunctive relief, see Lemley, *Licensing Market*, *supra* note 125, at \_\_\_. *Cf.* Alex Kozinski & Christopher Newman, *What's So Fair About Fair Use?*, 46 J. COPYRIGHT SOC'Y U.S.A. 513 (1999). But eliminating injunctive relief would still leave a punitive statutory damages regime in place.

151. We are skeptical that browserwrap terms of service giving companies like Google or Facebook plenary authority over what they do with your data are enforceable. *See* Mark A. Lemley,

A third, and less obvious reason, is that providing ML systems with broader access to data actually helps to mitigate some of the very negative outcomes that critics of ML systems fear. As the Obama White House recently identified, “AI needs good data. If the data is incomplete or biased, AI can exacerbate problems of bias.”<sup>152</sup> Facial recognition provides a good example. Facial recognition software performs worse at distinguishing individuals in small racial groups because since those groups are small, it has fewer unique data points allowing it to draw fine distinctions between faces in those groups. The solution is to build bigger databases overall or to “oversample” members of smaller groups. Ironically, this was exactly the motivation at play when two facial recognition tools—one by IBM and the other by MegaFace—came under fire in media outlets across the globe.<sup>153</sup> On

---

*Terms of Use*, 91 MINN. L. REV. 456 (2006). But if they are, they would create a blanket license for those companies to make different uses of the copyrighted material posted to their sites.

152. EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016), [https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/microsites/ostp/NSTC/preparing\\_for\\_the\\_future\\_of\\_ai.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf).

153. See, e.g., Kashmir Hill and Aaron Krolik, *How Photos of Your Kids Are Powering Surveillance Technology*, N.Y. TIMES (Oct. 11, 2019), <https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html> (noting: “Millions of Flickr images were sucked into a database called MegaFace. Now some of those faces may have the ability to sue.”); Olivia Solon, *Facial Recognition’s ‘Dirty Little*

202x]

Fair Learning clean 92620

155

both such occasions, the engineers involved collected millions of images from users that had released their photos under the creative commons license, which allowed for their bulk collection.<sup>154</sup> Both had resorted to collecting images outside of their internal datasets in hopes that exposure to a more diverse set of photos would help to increase the accuracy and reduce the potential for bias in their systems.<sup>155</sup> Yet, instead of being lauded, their efforts actually drew a swift rebuke from the media.<sup>156</sup>

---

*Secret': Millions of Online Photos Scraped Without Consent*, NBC NEWS (Mar. 17, 2019), <https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921>. Madhumita Murgia, *Who's Using Your Face? The Ugly Truth About Facial Recognition*, FIN. TIMES (2019), <https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e>.

154. Ryan Merkle, *Use and Fair Use: Statement on shared images in facial recognition AI* CREATIVE COMMONS (Mar. 13, 2019), <https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/>; *Artificial Intelligence and CC Licenses*, CREATIVE COMMONS (2019), <https://creativecommons.org/faq/#artificial-intelligence-and-cc-licenses>.

155. To draw accurate distinctions among faces, facial recognition software needs to train on lots of similar faces. Because large racial groups (or groups with heavy representation in a photo dataset) have more members than small ones, facial recognition software frequently performs less well in distinguishing members of racial minorities. Levendowski, *supra* note 32, at 589.

156. The bias objection and the privacy objection are at odds here. Many people worry about facial recognition because it will make mistakes. Correcting bias and getting better facial recognition helps alleviate that worry. But others may worry about facial recognition precisely because it *doesn't* make mistakes.

Many such articles deriding IBM and MegaFace also discussed IP rights as a potential mechanism to rein in these uses.<sup>157</sup> And though we're sympathetic to those fearful of the potential of ML systems to perpetuate bias, enable both private and public sector overreach, and run roughshod over some of society's most vulnerable citizens, trying to solve these problems by simply restricting access to more data is not a viable solution. As Amanda Levendowski convincingly argues, a permissive interpretation of fair use is "quite literally, [necessary for] promoting fairer AI."<sup>158</sup> In her telling, "The normative values embedded in the tradition of fair use align ultimately with the goal of mitigating bias. Fair use can, quite literally, promote creation of fairer AI systems."<sup>159</sup>

Finally, perhaps the strongest argument for fair use is one that lies at the heart of copyright theory but doesn't actually show up explicitly in the case law: Like TDM technologies, ML systems generally don't want to copy the copyrighted work for any copyright-related reason. ML systems generally copy works, not to get access to their creative expression (the part of the work the law protects), but to get access to the uncopyrightable parts of the work – the ideas, facts, and linguistic structure of the works. A self-driving car, for

---

157. See, e.g., *supra* note 153 and accompanying text.

158. See Levendowski, *supra* note 32, at 589.

159. See *id.* at 630.



instance, doesn't care about the composition or lighting of your photograph, or indeed about what you were likely actually intending to depict in your photo. It cares about the fact that there's a stop sign in it. Mapping software doesn't care what color you chose for your roads and political subdivisions; it wants to know where the roads are, what they are called, and which ones are one-way. Facial recognition software doesn't care about the composition choices (if any) you made in taking that selfie. It wants to know what you look like in a variety of lights, with and without a hat or a beard, and with different facial expressions. Search engines want to know what's in your work so they can help people find it, not because they care how you express yourself. Even ML systems that parse text or music usually don't care about the things that make those works copyrightable. They may be interested in corpus linguistics – how words are used in relationship to each other.<sup>160</sup> They may be training to understand or create natural language sentences by seeing how grammar is employed in practice. While there are some ML systems that train on art or writing in order to be able to create their own works of art, as discussed above, most are interested in copyrighted works for reasons that have nothing to do with the things that make those works copyrightable.

---

160. For discussion of the role of corpus linguistics in legal interpretation, see, e.g., Lawrence M. Solum, *The New Textualists' New Text*, 38 LOYOLA OF L.A. LAW REV. 2027 (date); Stephen C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU LAW REVIEW \_\_.

Ideas, facts, functions, methods, and stock literary and plot devices (aka scenes a faire) are not protectable by copyright law.<sup>161</sup> The “idea-expression dichotomy” (along with its cousins the fact-expression dichotomy and the process-expression dichotomy)<sup>162</sup> is perhaps the central doctrine in all of copyright law. Indeed, copyright courts have made it clear that copyright law wouldn’t be constitutional if it gave control over ideas and facts.<sup>163</sup> ML systems aren’t interested in expression (at least not for expression-related reasons); they just want the facts.

For humans, getting and using the unprotectable parts of a copyrighted work is normally not a problem. Read the book, or watch the movie, and you are free to take the ideas— or its standard plot elements — from that work for your own use. If it’s a factual work, you can use and indeed copy the facts in it without infringing the copyright. If it’s a computer program, you can take the functional aspects of that program, even if it means copying some of the

---

161. 17 U.S.C. § 102(b); *Baker v. Selden*, 101 U.S. 99 (1879) (establishing this precedent); Pamela Samuelson, *Why Copyright Law Excludes Systems and Processes from the Scope of Its Protection*, 85 TEXAS L. REV. 1921 (2007) (discussing the various things 102(b) refuses to protect).

162. *See* Samuelson, *supra* note 161.

163. *See* *Feist Publications, Inc., supra* note 71; *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539 (1985); *Eldred v. Ashcroft*, 537 U.S. 186 (2003).

202x]

Fair Learning clean 92620

159

code directly.<sup>164</sup> You can't copy the expression, but you don't need to in order to get access to the unprotectable elements. And once you have that access, you can reuse those elements without fear of liability.

That reflects an important, but rarely articulated, limit on the scope of copyright law. Unlike a patent, which gives its owner control over any "use" of the patented invention, a copyright only controls certain uses: copying, distributing, publicly performing, and the like. Notably absent from that list are certain activities fundamental to learning, such as watching, reading, and discussing a work and communicating its unprotectable elements to others.<sup>165</sup> Further, the first sale doctrine permits reselling, loaning, and otherwise sharing a single copy of a work.<sup>166</sup> These are all "uses" of the copyrighted

---

164. Well, except in the Federal Circuit. See *Computer Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693 (2d Cir. 1992); *Lotus Dev. Corp. v. Borland Int'l*, 49 F.3d 807 (1<sup>st</sup> Cir. 1995); Mark A. Lemley, *Convergence in the Law of Software Copyright?*, 10 BERKELEY TECH. L.J. 1 (1995) (discussing the consensus in the circuits on this point); Peter S. Menell, *Rise of the Copyright API Dead? An Updated Epitaph for Copyright Protection of Network and Functional Features of Computer Software*, 32 HARV. J. L. & TECH. 305 (2018); Pamela Samuelson & Clark D. Asay, *Saving Software's Fair Use Future*, 32 HARV. J. L. & TECH. (2018). *Contra Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339 (Fed. Cir. 2014) (holding such copying illegal), *cert. granted*, \_\_\_ S. Ct. \_\_\_ (Nov. 15, 2019).

165. See Jessica Litman, *The Exclusive Right to Read*, 13 CARDOZO ARTS & ENT. L.J. 29 (1994).

166. 17 U.S.C. § 109(a).

work, and important ones at that. But they aren't uses the law forbids, because the point of copyright is not just to give incentives to create but to "promote the progress of science"<sup>167</sup> by ensuring those creations are shared by others. The freedoms to read, to learn, and to communicate what you have learned are critical to making the idea-expression dichotomy work in practice, because it helps ensure people can find the ideas in a copyrighted work in order to use them.

True, there are some circumstances in the pre-ML world in which giving copyright owners control over expression may risk allowing them to lock up the unprotectable ideas and facts as well. That is particularly likely when the work is predominantly composed of uncopyrightable elements. In those circumstances, courts and Congress have created a variety of doctrines to help ensure the user's access to the unprotectable parts of the copyrighted work. We impose a higher standard for proof of infringement of factual works whose copyright is "thin," requiring "virtual identity" of the works rather than merely "substantial similarity."<sup>168</sup> Doing so reduces the risk that a defendant will wrongly be held liable if their work is too similar to the

---

167. U.S. CONST. art. I, § 8, cl. 8.

168. *See, e.g.,* *Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972 (2d Cir. 1980). The Ninth Circuit questioned whether this was a different standard in its en banc decision in *Skidmore v. Led Zeppelin*, \_\_\_ F.3d \_\_\_ (9<sup>th</sup> Cir. 2020) (en banc), but it upheld the virtual identity requirement.

202x]

Fair Learning clean 92620

161

plaintiff's because it shares the uncopyrightable facts and ideas of the plaintiff's work. We theoretically give a broader scope of fair use to defendants copying factual works, though some courts deny that doctrine has much if any force.<sup>169</sup> We deny protection to copyrightable expression altogether if the work is so short<sup>170</sup> or so bound up with the ideas that there are only a limited number of ways of expressing those ideas. In that case we say the idea and expression have "merged" and refuse to protect the expression for fear of also locking up the ideas.<sup>171</sup> We are supposed to do the same for "useful articles" where we can't separate the creative from the

---

169. See 17 U.S.C. §106(2). Compare *Oracle Corp. v. Google Inc.*, 886 F.3d 1179 (Fed. Cir. 2018), *cert. granted* Nov. 15, 2019) and *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169 (2d Cir. 2018) (both treating this factor as insignificant) with *Am. Soc'y for Testing & Materials, et al. v. Public.Resource.Org, Inc.*, 896 F.3d 437 (D.C. Cir. 2018); (treating it as central)."

170. The Copyright Office's circular, for example, notes that "slogans, and other short phrases or expressions cannot be copyrighted." Though not having the force of a statute, it is considered to be "a fair summary of the law." *Kitchens of Sara Lee, Inc. v. Nifty Foods Corp.*, 266 F.2d 541 (2d Cir. 1959); *Ets-Hokin v. Sky Spirits, Inc.*, 225 F.3d 1068 (9th Cir. 2003). *But see, e.g.*, *CRA Mktg., Inc. v. Brandow's Fairway Chrysler-Plymouth-Jeep-Eagle, Inc.*, 1999 U.S. Dist. LEXIS 11889 (E.D. Pa. 1999) (finding 54 words was enough); *Hall v. Swift*, \_\_\_ F.3d \_\_\_ (9th Cir. Oct. 28, 2019) (holding that Taylor Swift could infringe copyright in a song with only six similar words: "playas gonna play, haters gonna hate").

171. See, e.g., *Morrissey v. Proctor & Gamble Co.* 379 F.2d 675 (1st Cir. 1967); *Sampson & Murdock Co. v. Seaver-Radford Co.*, 140 F.539, 541 (1st Cir. 1905); cf. KAPLAN, AN UNHURRIED VIEW OF COPYRIGHT, 64-65 (1967).

functional aspects,<sup>172</sup> though the Supreme Court recently muddied the waters around that doctrine to such an extent that it's no longer clear we deny protection in such a case.<sup>173</sup>

All these doctrines give less, or in some cases no, protection to a copyrighted work because the ideas and facts are so closely bound up with the expression in that work that it is hard if not impossible for users to take the unprotectable bits without treading on the protectable ones. They focus, in other words, on the centrality of ideas to the plaintiff's work and the difficulty in separating the protectable from the unprotectable.<sup>174</sup> And they limit protection or deny it entirely to works that are primarily composed of uncopyrightable matter.

---

172. See 17 U.S.C. § 101.

173. See *Star Athletica, L.L.C. v. Varsity Brands, Inc.*, 137 S. Ct. 1002, 197 L. Ed. 2d 354 (2017). For a small sample of the numerous trenchant criticisms of that case, see Christopher Buccafusco, Mark A. Lemley & Jonathan S. Masur, *Intelligent Design*, 68 DUKE L.J. 75, 121-23 (2018); Mark P. McKenna, *Knowing Separability When We See It*, 166 U. PA. L. REV. ONLINE 127 (2017); Jennifer Yamin, *Interview: Professor Fromer and the Star Athletica Case*, NYU LAW BLOG (Apr. 5, 2018), <https://blog.jipel.law.nyu.edu/2018/04/interview-professor-fromer-and-the-star-athletica-case/>; Barton Beebe, *Star Athletica and the Problem of Panaestheticism*, 9 UC IRVINE L. REV. 275 (2019).

174. See Mark A. Lemley & Philip J. Weiser, *Should Property or Liability Rules Govern Information?*, 85 TEXAS L. REV. 783 (2007).

But that won't help our robotic learners. They aren't taking works that are particularly factual or functional in nature. Some might be – protection for photographs is or ought to be thin, for instance<sup>175</sup> – but some will be ordinary copyrighted works. The problem ML systems face is the inability to capture the unprotectable parts to use for training without making a rote copy of the protectable ones. Systems want access to the unprotectable bits of creative works, but the way they get that access is necessarily copying the whole thing.<sup>176</sup> Unlike humans, they can't read to learn or observe the idea in a painting or song without making a copy of the whole thing in their training data set.

#### D. *Fair Machine Learning*

Above, we saw a host of potential legal challenges await those seeking to build training sets for ML systems and that there are compelling reasons for why policymakers should be concerned. In this subpart, we suggest that the analysis of fair use for AI training data should incorporate a principle we call “fair learning.” If the purpose of the AI's use is not to obtain or

---

175. See, e.g., Jessica Silbey, *Justifying Copyright in the Age of Digital Reproduction: The Case of Photographers*, 9 UC IRVINE L. REV. 405 (2019); Shyamkrishna Balganesh, *Causing Copyright*, 117 COLUM. L. REV. (2017).

176. See Grimmelman, *supra* note 18 (describing the technical architecture that makes copying mandatory).

incorporate the copyrightable elements of a work but to access, learn, and use the unprotectable parts of the work, that use should be presumptively fair under the first fair use factor (the purpose of the use). Notably, fair learning by ML systems should be fair even if fair use factors two and three (the nature of the work and the amount taken) would otherwise weigh against fair use. Systems should be able to learn from fictional as well as factual works,<sup>177</sup> and

---

177. The second factor also gives special protection to unpublished works in order to preserve the plaintiff's right of first publication. *Harper & Row v. Nation Ents.*, 471 U.S. 539 (1985). But that shouldn't apply here because the AI isn't publishing the plaintiff's work at all and therefore isn't preempting the right of first publication. Some plaintiffs use copyright in order to hide facts from the world. *Salinger v. Random House*, 811 F.2d 90 (2d Cir. 1987) (enjoining publication of J.D. Salinger's letters); *Shloss v. The Estate of James Joyce*, \_\_\_ F. Supp. 2d \_\_\_ (N.D. Cal.) (refusing effort by James Joyce's heirs to stop publication of his letters); *Religious Technology Center v. Lerma*, 908 F. Supp. 1362 (E.D. Va. 1995) (Church of Scientology suing to prevent publication of its scriptures); *Houghton Mifflin Co. v. Stackpole Sons, Inc.*, 104 F.2d 306 (2d Cir. 1939) (enforcing Adolf Hitler's copyright claim against Alan Cranston, who sought to translate the full version of *Mein Kampf* to show Hitler's true beliefs). But those are not socially desirable uses of copyright. After *Salinger*, Congress amended section 107 to prevent the unpublished nature of the work from being conclusive against a finding of fair use. See Daniel E. Wanat, *Fair Use and the 1992 Amendment to Section 107 of the 1976 Copyright Act: Its History and an Analysis of Its Effect*, 1 VILL. SPORTS & ENT. L.F. 47 (1994) (noting that the legislative history of the amendment "reflects an intention to remedy the perceived chilling effect of *Salinger*"). And we think the argument for fair use is stronger, not weaker, if the plaintiff uses copyright not to control the timing and profit of first publication but to prevent publication of facts altogether.



202x]

Fair Learning clean 92620

165

ML systems naturally learn by reviewing (and therefore “taking”) the entire work. The fourth factor (market effect) should normally not prevent fair learning use of individual copyrighted works. The copyright owner of a book or photograph doesn’t create that work in hopes of selling it to AIs. It is possible that they might make some additional money from licensing the work to AIs. But the mere existence of a licensing market directed to such uses shouldn’t make it unfair, just as a “licensing market” for radically transformative uses doesn’t defeat the value of transformation.<sup>178</sup> Only if the use directly interferes with the plaintiff’s core market should the fourth factor outweigh a finding of fair learning under the first factor.

That doesn’t mean that it should always be fair use for ML systems to copy data for use in a training set. The purpose to which the ML system ultimately puts the information may matter to several of the fair use factors. Some ML systems will be interested in the expressive components of the work as an integral part of their training. That is, the goal will be to teach the system using the creative aspects of the work that copyright values, not just using the facts or the semantic connections the law is not supposed to protect. That is particularly likely of those systems like MuseNet that are training in order to generate their own expressive works. Those ML systems both copy

---

178. *See Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605 (2d Cir. 2006).

expression for expression's sake and pose a threat of "significant substitutive competition" to the work originally copied.<sup>179</sup>

Learning by such systems might still be fair. First, we might distinguish between the input and the output, and say that the act of learning itself should be protected even if you learn from copyrighted work, just as humans can learn music by singing songs or learn literature by reciting poems. So perhaps an AI that learns what makes an Ariana Grande song an Ariana Grande song should be free to do so even though it cares about the expression, not just the facts, just as humans could.<sup>180</sup> Certainly an AI that creates parodies of a song should be entitled to the same fair use protection a human would.<sup>181</sup>

The problem comes when we ask what we want such an AI to do with that information. What is the output of that AI? Some answers won't be

---

179. See Sag, *The New Legal Landscape for Text Mining and Machine Learning*, *supra* note     (arguing that AI copying to create new works is a tougher case for fair use); Benjamin L.W. Sobel, *Elements of Style: Emerging Technologies and Copyright's Fickle Similarity Standards* at 57-58 (working paper 2020) (arguing that AI that produces new art in the style of an existing artist likely "would be unable to avail itself of the fair use defense").

180. We are indebted to James Grimmelmann for this point.

181 *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994) (finding parody to qualify as fair use). In fact, an AI named "Weird A.I. Yancovic" that writes song parodies has had its songs taken down by the recording industry. Katie Canales, *A researcher created a 'Weird A.I. Yancovic' algorithm that generates parodies of existing songs, and now the record industry is accusing him of copyright violations*, BUS. INSIDER, July 24, 2020.

worrisome from a copyright perspective. We might train an AI to recognize an Ariana Grande-like song in order to try to catch infringers of her songs, for instance. More likely, the AI will produce creative works as its output. Even that isn't necessarily unfair. Many of the works created by systems like MuseNet will be transformative uses that society values. But it makes the fair use case closer, because the output of the ML's learning competes with the plaintiff's core market. And some purposes – say, a system designed to write a new pop song in the style of Taylor Swift or a translation program that produces a translation of an entire copyrighted work – seem more substitutive than transformative,<sup>182</sup> so that if they run afoul of the ever-broadening definition of similarity in music,<sup>183</sup> fair use is unlikely to save them.<sup>184</sup>

---

182. For a discussion of “style appropriation” by AI and its copyright implications, see Sobel, *style*, *supra* note 179.

183. See, e.g., *Williams v. Gaye*,—F.3d—(9<sup>th</sup> Cir. 2018); *Hall v. Swift*,—F.3d—(9<sup>th</sup> Cir. 2019) (permitting suit against Taylor Swift based on six-word phrase “players gonna play, haters gonna hate”); Sobel, *Style*, *supra* note 179 (discussing the “fickle” treatment of similarity across copyright’s different domains). *But see* *Skidmore v. Led Zeppelin*, — F.3d — (9<sup>th</sup> Cir. 2020) (en banc) (rejecting copyright claim based on similarity of simple musical phrases).

<sup>184</sup> Gilotte argues that there is unlikely to be market harm because the risk of market substitution is small for those customers who want a known artist’s work. Gilotte, *supra* note \_\_\_, at 2688. While that is true, AI-generated works may displace sales by other, lesser-known artists who are interested in the appearance or sound of the work but not in the brand name.

Fair learning will also properly distinguish between the use of individual copyrighted works in a training dataset – the issue we are primarily concerned with here – and the wholesale copying of a competitor’s training dataset. Copying a copyrighted database generally involves copying the things copyright does protect: the selection and arrangement of the data. So the purpose of the use won’t normally favor the copier in such a case.<sup>185</sup> Even if it did, taking an entire training database is likely to have a direct market effect, since the value of that database, unlike the value of any individual copyrighted work, is in its use for ML training. Systems are using copyrighted works, but, generally, they’re not using them in ways inimical to copyright’s purposes. Fair learning properly takes that fact into account.<sup>186</sup>

Other scholars have focused on the nature of the entity doing the copying<sup>187</sup> or whether the use being made by the defendant is itself

---

185. There may, however, be cases in which the defendant is uninterested in the things that makes the database copyrightable and only interested in the unprotectable facts. For instance, if I don’t want your curated database of representative faces, but simply want to collect as many faces as possible for my dataset, I am arguably not copying your database for the purpose of taking the selection and arrangement that make your database copyrightable in the first place.

<sup>186</sup> For a parallel argument under European law, see Mauritz Kop, *The Right to Process Data for Machine Learning Purposes in the EU* (working paper 2020).

187. See, e.g., Grimmelmann, *supra* note 18.

202x]

Fair Learning clean 92620

169

communicating to the outside world rather than an internal use.<sup>188</sup> Both of those approaches overlap with ours, but they aren't focused on the key question. Fair learning isn't fair because it is a machine doing it, or because it happens outside the public view. It's fair because the value the ML system gets from the copyrighted work stems from the part of the work the copyright law has decided belongs to the public, not to the copyright owner.<sup>189</sup>

### III. Fair Learning Beyond AIs

#### A. *Copyright for Literate Humans*<sup>190</sup>

It might seem unfair that humans have to pay for copyrighted works when AIs (often owned by giants like Google) don't.<sup>191</sup> But a fair learning doctrine won't just benefit machines. Humans want to learn too. As we saw in section II.C, the law generally lets them learn without fear of copyright

---

188. Sag, *supra* note 48, at 18 (“Non-expressive use is also justified in terms of expressive substitution, but even more emphatically so. By definition, a non-expressive use does not usurp the copyright owner’s communication of her original expression to the public because the expression is not communicated.”).

189. Grimmelmann doesn't take this position directly, but he might well agree with us.

190. With apologies to James Grimmelmann. Grimmelmann, *supra* note 18.

191. The AI will likely have to get lawful access to the work somehow, so they will often have to pay for works that the copyright owner didn't make freely available online. Fair learning protects the act of copying; it may not convey a right to access a work not otherwise publicly available.

liability. And when that learning is at risk, we've created several doctrines to make sure that humans have access to the unprotectable aspects of a copyrighted work. But those doctrines don't always work. Fair learning won't just protect robots. It might apply to human learning too in a variety of situations where that learning is at risk. Understanding how fair use protects ML may also help courts do a better job of identifying and protecting fair learning by humans too.

First, the right to read a book or watch a show isn't guaranteed in the statute. It has historically been a function of the physical way in which content was embodied. You don't need to make a copy of a book in order to read it, to sell it at a used bookstore, or to lend it out at a library. So none of those things traditionally implicated copyright law. But as consumption of content has moved from physical media to computers, that has changed. Essentially everything anyone can do with a copyrighted work on a computer involves making one (and usually many) copies.<sup>192</sup> That means that while you might have the right to read or resell your physical book, you don't have the same rights with your ebook.<sup>193</sup> Indeed, some courts have held that even

---

192. See Mark A. Lemley, *Dealing with Overlapping Copyrights on the Internet*, 22 U. DAYTON L. REV. 547 (1997).

193. See Chris Jay Hoofnagle et. al., *The Tethered Economy*, 87 GEO. WASH. L. REV. 783 (2019); Litman, *supra* note 165.

202x]

Fair Learning clean 92620

171

turning on your computer is illegal because it copies the operating system software into short-term memory.<sup>194</sup> Companies that have tried to create a virtual equivalent of reselling a used book have been sued for copyright infringement and lost.<sup>195</sup> Humans today, in other words, face a similar problem to ML technologies— they have to copy the whole digital work even if they only want to learn the facts or ideas in it. Fair learning can ensure that humans too don't lose the right to read simply because the book they are reading has moved online.

Computers aren't the only circumstance in which copyright law might punish humans who just want the facts from a work. In *American Geophysical Union v. Texaco*,<sup>196</sup> the Second Circuit held that research scientists did not engage in fair use when they made photocopies of scientific journal articles circulated by the in-house library for a research file or to take into the lab with them.<sup>197</sup> The case has occasioned much commentary and criticism, most of it focused on the court's expansive use of a licensing

---

194. MAI Sys. Corp., 991 F.2d 511. Congress overruled this case, but only narrowly, providing a specific carveout allowing people to turn on computers in order to diagnose and repair them. 17 U.S.C. § 117.

195. Capitol Records, LLC v. ReDigi Inc., 910 F.3d 649 (2d Cir. 2018), *cert. denied*, 139 S. Ct. 2760 (2019).

196. *Texaco, Inc.*, 60 F.3d 913, 930 (2d Cir. 1994).

197. *Id.* at \_\_.

market to sway the fourth factor against *Texaco*.<sup>198</sup> But we can also see *Texaco* as a fair learning case. The scientists who made the photocopies didn't want the articles in the *Journal of Catalysis* for their beautiful prose style or even the clever structure, sequence, and organization of the paragraphs. They wanted access to the facts in the article: the graphs and data results. Photocopying the article was a convenient alternative to writing the results down (something that clearly would have been legal). But photocopying, like reading on a computer, makes a copy of the whole work, not just the facts, so the court held it illegal. Because the scientists weren't commercializing the journal articles and weren't interested in the things that made those articles copyrightable, *Texaco* seems a plausible candidate for fair dealing in the context of automated reproduction – less automated than ML training, to be sure, but automated still.<sup>199</sup>

Fair learning may also help resolve the current dispute over the copyright status of the law itself. This may seem an odd topic to be controversial. Surely the law itself is free for the public to use? Federal law

---

198. See, e.g., Lemley, *Licensing Market*, *supra* note 125; Ann Bartow, *Educational Fair Use in Copyright: Reclaiming the Right to Photocopy Freely*, 60 U. PITT. L. REV. 149, 207-21 (1998). Samuelson, *Unbundling*, *supra* note 217 (*Texaco* has “caused a good deal of agitation and anxiety in educational, library, and research communities”).

199. *Texaco*, 60 F.3d 913, 930



202x]

Fair Learning clean 92620

173

is.<sup>200</sup> But the status of official court reporters, official annotations to state codes, local ordinances like planning maps, and private industry codes and standards that are adopted into law is less clear. Copyright owners have filed suit over each category of work. While courts have so far mostly rejected those suits,<sup>201</sup> they have struggled to articulate a theory for why a new standard or ordinance couldn't be copyrighted and, if it is, why wholesale copying should be permitted. Some courts have held the ordinances unprotectable under merger,<sup>202</sup> while others have held the copying of those ordinances to be fair use.<sup>203</sup> In 2020 the Supreme Court held that the annotations to state statutes were not protected by copyright because they were "government edicts" given legal significance even though drafted by private parties.<sup>204</sup> But it left open the possibility that other documents adopted into law, such as private standards required by regulation, might retain copyright protection. If the Court does find them copyrightable, we think

---

200. 17 U.S.C. §105.

201. *West Pub. Co. v. Mead Data Cent., Inc.*, 616 F. Supp. 1571 (D. Minn. 1985); *Veeck v. Southern Bldg. Code Congress Int'l, Inc.*, 293 F.3d 791 (5th Cir. 2002) (en banc); *Am. Soc'y for Testing & Materials, et al. v. Public.Resource.Org, Inc.*, 896 F.3d 437 (D.C. Cir. 2018); *Georgia v. Public.Resource.Org, Inc.*, 140 S.Ct. 1498 (2020). Full disclosure: one of us (Lemley) represents Public Resource in these cases.

202. *Veeck*, 293 F.3d at 791; *Georgia*, 140 S.Ct. at 1498.

203. *Am. Soc'y for Testing & Materials*, 896 F.3d at 437.

204. *Georgia*, 140 S.Ct. at 1498.

even wholesale copying of industry standards mandated by regulation for any purpose other than commercial sale is likely to be protected by fair learning. People don't read statutes or regulations for entertainment; they read them for the legal mandates they contain. Learning what law governs a person's behavior seems a particularly important form of learning in a democratic society.

We might also apply fair learning to excuse copying of newsworthy material for the purpose of reporting the news. That would require changing some case law, particularly the cases involving videos of famous historical events. The person who happens to take a video of the Kennedy assassination or the Reginald Denny beating has created a copyrighted work,<sup>205</sup> but viewers and the news media aren't interested in the copyrighted bits. They want to see what actually happened, not the accidents of the plaintiff's angles and lighting. Because often only one person filmed the event, anyone who wants to watch that event needs to watch the copyright owner's version. Giving control to the copyright owner locks up the unprotectable as well as the protectable parts, and here those are the parts that people care about. Nonetheless, most, though not all, cases involving media reporting of news

---

205. Gorman, *Copyright Protection for the Collection and Representation of Facts*, 76 HARV. L. REV. 1569, 1597-1598 (1963).

202x]

Fair Learning clean 92620

175

videos have rejected fair use claims.<sup>206</sup> We think that is wrong. Selling the video for profit might not be fair use, but using it to learn what happened should be.

We might extend the concept of fair learning even further to the analogous idea of “fair functioning.” Many copyrighted things serve a purpose beyond just communication.<sup>207</sup> That list includes software, which is a literary work (code is written) but is for all practical purposes a functioning machine;<sup>208</sup> a sequence of yoga poses, which can be creative and expressive but which are clearly designed to train the body and not just as a form of

---

206. *Monge v. Maya Magazines, Inc.*, 688 F.3d 1164, 1176 (9th Cir. 2012) (photos of celebrity couple’s secret wedding “did not transform the photos into a new work . . . or incorporate the photos as part of a broader work”). *L.A. News Serv. v. Reuters Television Int’l*, 149 F.3d 987 (9th Cir. 1998) (holding that a rebroadcast of a video clip of beating of Reginald Denny during the 1992 Los Angeles riots was not fair use), *and L.A. News Serv. v. KCAL-TV Channel 9*, 108 F.3d 1119 (9th Cir. 1997) (same), *with L.A. News Serv. v. CBS Broad., Inc.*, 313 F.3d 1093 (9th Cir. 2002) (holding that a shorter use of the same clip was fair use); *Time v. Bernard Geis Assocs.*, 293 F. Supp. 130 (S.D.N.Y. 1968) (finding the use of frames from the Zapruder film of the Kennedy assassination to be fair use); *Núñez v. Caribbean Int’l News Corp.*, 235 F.3d 18 (1st Cir. 2000).

207. *Cf.* Michael J. Madison, *Law as Design: Objects, Concepts, and Digital Things*, 56 CASE W. RES. L. REV. 381 (2005);

208. Dan L. Burk, *Method and Madness in Copyright Law*, 2007 UTAH L. REV. 587; Samuelson, *Unbundling Fair Uses*, *supra* note 217.

dance;<sup>209</sup> and clothing design, which combines both artistry and function.<sup>210</sup> Just as the desire to access the facts and ideas of a written work should influence the analysis of fair use, arguably the defendant's desire to make use of the functional aspects of a useful work rather than its artistic ones – to practice Bikram-style yoga, to make a computer program work with another, and so on – should be a factor that favors a finding of fair use.

Fair learning, in short, isn't just for machines. Humans too may need a practical, fair use-based right to learn and to do the things that copyright law nominally says they can learn and do.

*B. Toward a Pluralist Theory of Fair Use*

Treating fair learning as a lawful purpose under the first factor also offers some broader lessons for fair use doctrine more generally. It provides a desirable counterbalance to the recent emphasis on transformative use, opening the way to a more pluralistic vision of fair use. Transformative use has arguably swallowed fair use doctrine in the past 25 years.<sup>211</sup> Transformation is important. Transformative works are themselves creative works that copyright law should encourage, not discourage. But the rush to

---

209. *Bikram's Yoga College of India LP v. Evolution Yoga LLC*, \_\_ F.3d \_\_ (9<sup>th</sup> Cir. 2015).

210. *Star Athletica, L.L.C. v. Varsity Brands, Inc.*, 137 S. Ct. 1002, 197 L. Ed. 2d 354 (2017).  
*See also supra* note 173 for a host of critiques of that decision.

211. *See supra* notes 92–97 and accompanying text.

make transformative use the centerpiece of fair use doctrine has obscured the fact that uses need not be transformative to be fair.<sup>212</sup> Some of the classic examples of fair use, such recording a song or television show at home, are fair not because the defendant did anything new or creative but because they aren't commercial and don't have any likely market effect.<sup>213</sup> Other examples, like copies for classroom use<sup>214</sup> or the reproduction of images from the Zapruder film of the Kennedy assassination,<sup>215</sup> are fair not because they are transformative or because they have no market consequence but because they serve valuable social purposes, educating students or permitting informed discussion of political and social issues.<sup>216</sup> Fair use has long been

---

212. Tony Reese emphasizes that use of a work might have a transformative purpose even if it doesn't transform the content of the work. R. Anthony Reese, *Transformativeness and the Derivative Work Right*, 31 *COLUM. J.L. & ARTS* 101 (2008). Michael Carroll argues that compiling a database for research purposes *is* a transformative purpose, though not a transformation of the content. Carroll, *supra* note 106, at 941-44.

213. *Sony*, 203 F.3d 596.

214. 17 U.S.C. § 107.

215. *Time Inc. v. Bernard Geis Assocs.*, 293 F. Supp. 130 (S.D.N.Y. 1968).

216. Adolf Hitler's suit against Alan Cranston for translating the full version of *Mein Kampf* to show how the official translation had been sanitized was held not to be a fair use, *Houghton Mifflin Co. v. Stackpole Sons*, 104 F.2d 306 (2d Cir. 1939), but it seems a quintessential example of a public benefit that should have been held fair. A number of scholars have recently challenged the primacy of transformative use. Haochen Sun argues that the public interest factor should play a larger role in fair use regardless of whether the use was transformative. Haochen Sun, *Copyright Law as an*

about more than transformation. It is important to recall that more pluralistic vision of fair use.<sup>217</sup>

Fair learning adds two important policy rationales to this pluralistic vision of fair use. First, it addresses the problem of overinclusiveness in copyright enforcement. Copyright law gives owners control over some parts of their work but not others. When users can't separate the protectable from the unprotectable parts, however, control over part can easily become control over all. That can happen as an accident, but it can also happen deliberately. Copyright owners regularly use the law as a tool to prevent disruptive competition that threatens their incumbent markets.<sup>218</sup> That's why the law denies protection altogether in the merger and inseparable useful article cases. But denying all protection is unfair to copyright owners in many cases because they have in fact contributed substantial expression the law wants to encourage. Another approach is to limit remedies so that they are consonant with the scope of what is actually protected.<sup>219</sup> That may mean denying injunctive relief and reforming damages in automated copying cases, as

---

*Engine of Public Interest Protection*, 16 NW. J. TECH. & INTELL. PROP. 123 (2019). See also Asay, *supra* note 8; Alexander McMullan, *Returning to the Fair Use Standard*, 63 N.Y.L.S. L. Rev. 329 (2018-19). We agree with those critiques.

217. Pamela Samuelson, *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537 (2009).

218. Lemley & McKenna, *Unfair Disruption*, *supra* note 121.

219. Mark A. Lemley & Mark P. McKenna, *Scope*, 57 WM. & MARY L. REV. 2197 (2016).

Lemley and Weiser have proposed.<sup>220</sup> But while that is achievable when it comes to injunctions,<sup>221</sup> it would require changes to the damages statute. Treating fair learning as fair use can help to calibrate the scope of copyrights, ensuring that copyright owners get control over expressive elements and uses of their work but can't leverage that right to effectively control the unprotectable elements of their work.<sup>222</sup>

Fair learning offers a second theoretical lesson, one that goes to the heart of the purpose behind copyright law. A central problem with allowing copyright suits against ML is that the value and benefit of the system's use is generally unrelated to the purpose of copyright. That is true not only because the ML system wants the facts, ideas, and other unprotectable elements of the work. Arguably it's true even if the technology gains a non-expressive benefit from the expressive parts of the work too, perhaps by learning how to recognize a particular artist's song or painting. That is use of the protectable expressive parts of the plaintiff's work. But the ML system doesn't care whether the work is expressive or not and which aspects are protected. It just wants to learn from the work in order to put that knowledge to a different

---

220. Lemley & Weiser, *supra* note 174, at \_\_\_.

221. *See* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006); *Salinger v. Colting*, 607 F.3d 68 (2d Cir. 2010).

222. Samuelson, *Unbundling Fair Uses*, *supra* note 217; Mark A. Lemley & Mark P. McKenna, *Scope*, 57 WM. & MARY L. REV. 2197 (2016).

instrumental use.<sup>223</sup> Perhaps this should be a broader principle of fair use, one not limited to fair learning: if the defendant has no interest in the work because of the thing that makes that work copyrightable, the use is presumptively not one that interferes with the purpose of copyright law, and so ought to be considered fair.

Whether or not you agree with us that fair learning by ML should be fair use, the concepts underlying fair learning are concepts fair use doctrine should take into account. ML systems therefore have much to teach us about copyright law for humans too. As James Grimmelman notes, “paying attention to robotic readership refocuses our attention on the really fundamental questions: what is copyright, and what is it for? To say that human readers count and robots don’t is to say something deep about the nature of reading as a social practice, and about what we want robots— and humans—to be.”<sup>224</sup> We don’t think the law should treat robots and humans differently. On the contrary, each should be entitled to learn from a copyrighted work in the way they naturally learn.

---

223. Yeah, we know, we’re anthropomorphizing AIs. They don’t really “want” anything. Deal with it. As we’ve noted elsewhere, everyone does it. *See* Casey & Lemley, *supra* note 7.

224. Grimmelman, *Copyright for Literate Robots*, *supra* note 18.



202x]

Fair Learning clean 92620

181

## Conclusion

Machine learning requires copying extraordinary amounts of copyrighted material. That copying should generally be permitted. Most ML systems copy works not in order to consume the expression copyright law protects, but to get access to the facts or structures copyright law dedicates to the public. Understanding this as fair learning can help ensure we can train ML systems without interference from the law. But the idea of fair learning doesn't just matter for robots. It can help us resolve a number of troubling copyright cases involving humans too. And it reminds us that fair use is about more than just transforming copyrighted works into new works. It's about preserving our ability to create, share, and build upon new ideas. In other words, it's about preserving the ability to learn – whether the entity doing the learning is a person or a robot.