Essay

Racial Bias in Algorithmic IP

Dan L. Burk[†]

> "Justice?" Hawkmoon called after him as he left the room. "Is there such a thing?"
>
> "It can be manufactured in small quantities," Fank told him. "But we have to work hard, fight well and use great wisdom to produce just a tiny amount."[1]

INTRODUCTION

Intellectual property law currently stands at the intersection of two dramatic social trends. Machine learning systems, a form of artificial intelligence, are increasingly being deployed across a range of social practices, including the development of innovative or creative works and the administration of intellectual property (IP) rights associated with those works.[2] At the same time, evidence of racial bias in IP systems is manifest and growing: as with many social practices, scholars and practitioners have begun to seriously contemplate the reform of intellectual property systems that have historically excluded disfavored minorities.[3]

The confluence of these trends has not gone unnoticed, and legal scholars have already begun to ask whether the biases present in existing IP systems may infect algorithmic processes trained on data

[†] Chancellor's Professor of Law, University of California, Irvine. My thanks to Mark Lemley, Orly Lobel, Brenda Simon, Felix Wu, and participants at the 2021 Works in Progress Intellectual Property Colloquium for their comments on a previous version of this paper. Copyright © 2022 by Dan L. Burk. 12

[1] MICHAEL MOORCOCK, THE SECRET OF THE RUNESTAFF 501 (1969).

[2] *See* Daniel Gervais, *Is Intellectual Property Law Ready for Artificial Intelligence?*, 69 GRUR INT'L 117 (2020).

[3] *See* Anjali Vats & Deidre A. Keller, *Critical Race IP*, 36 CARDOZO ARTS & ENT. L.J. 735 (2018) (calling for the development of a critical racial perspective in intellectual property law).

from past practices.[4] Although such critiques properly express concerns about the algorithmic entrenchment of bias, they say little about the mechanisms by which this may occur, and so are less helpful than we might hope in understanding what, if anything, might be done about problematic outcomes. It would be desirable to better explore the nature of the biases we might expect AIs to perpetuate in intellectual property systems, something of the mechanisms that might lead to bias in AI administered intellectual property, and the potential for entrenchment of bias in development or administration of intellectual property via AI systems.

Consequently, in this Essay I attempt to identify certain social bias problems that in the context of intellectual property will be particular to the algorithmic determinations through AI processing. I begin by describing the convergence of trends in intellectual property: the implementation of AI-driven systems alongside the recognition of longstanding racial disparity in IP. I then disambiguate some of the existing literature dealing with "bias" in AI, distinguishing discussions of technical bias from social, and more particularly racial, biases. This allows me to separate questions of accuracy from questions of social discrimination, and to show how proposals to correct the former are unlikely to correct the latter. Although our understanding of bias in IP is still nascent, vignettes from other areas where law and IP have intersected illuminate areas of concern for IP practice. Finally, I identify socially biased effects of AI systems that pose different challenges to intellectual property than past biases now being identified by scholarship on IP and race. I close with some observations on the use of AI as a diagnostic for intellectual property, rather than as a constitutive feature.

I. ARTIFICIAL INTELLIGENCE AND IP

We should perhaps begin with an observation that I have made in previous work, but which seems to bear repeating wherever artificial intelligence and the law is under consideration, which is that "artificial intelligence" is something of an unfortunate misnomer.[5] The

[4] *See, e.g.*, W. Keith Robinson, *Artificial Intelligence and Access to the Patent System*, 21 NEV. L. REV. 729 (2021); *see also* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018) (arguing that copyright law will both contribute to and ameliorate social bias in AI training data).

[5] *See* Dan L. Burk, *AI Patents and the Self-Assembling Machine*, 105 MINN. L. REV. HEADNOTES 301, 303 (2021).

technology under consideration entails nothing intelligent in any robust sense of that word—research has been done on such machine cognition, and continues, but there is no realistic prospect building of anything like human (or even animal) cognitive facility into a machine anywhere in the foreseeable future.[6] There is in fact good reason to believe that, absent a revolution in the available technology, such a goal is unachievable at all.

The systems that have garnered recent attention belong to a subset of artificial intelligence research with far narrower and more modest objectives.[7] These fall under the label of "machine learning," which is itself also a somewhat unfortunate nomenclature, again because of potentially misleading associations attached to the analogy to "learning": machines do not learn in the sense that most people would commonly use or understand that term. The technology might be better termed as "pattern recognition" systems—although, again, with some cautions about the use of the term "recognition." The most appropriate label for the technologies in question might be "statistical optimization" systems, because this is in fact what the technology does: leverages very fast processing power and cheap computer data storage to iteratively fit increasingly better statistical models to very large data sets.[8]

Labeling the technology as "statistical optimization" also avoids the unfortunate comparisons to human cognition, and the accompanying analytical distortions of anthropomorphizing the technology. Such comparisons are hyperbolic and unnecessary. Even without romanticizing their characteristics, modern machine learning systems are impressive in their ability to parse data sets that would otherwise be unmanageable, and to find correlations within such data that would remain hidden from unaided human scrutiny. These technical capabilities are advancing and are increasingly applicable to a large range of circumstances, including many applications that will generate products that fall within the scope of intellectual property law.

[6] *See* Madeline Clare Elish & danah boyd, *Situating Methods in the Magic of Big Data and AI*, 85 COMM. MONOGRAPHS 57, 61 (2017) (describing the failure of research into general artificial intelligence).

[7] *See* Marion Fourcade & Kieran Healy, *Seeing Like a Market*, 15 SOCIO-ECON. REV. 1, 24 (2017) (observing that AI research abandoned the idea of machines that can think in favor of machines that can learn).

[8] *See* Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1, 5–6 (2016) (explaining the statistical optimization modeling typical of machine learning systems).

For example, AI systems are expected to be employed in the future design, selection, and deployment of trademarks.[9] In the areas belonging to copyright subject matter, AI systems can be trained to generate works in a range of expressive or aesthetic fields, such as new graphic images in the style of past artists, for example resembling the art of Rembrandt.[10] They can similarly generate images that combine known styles or develop images in previously unknown styles.[11] The same is true for the generation of new musical compositions derived from data on past styles of music.[12] AIs can be used to develop new choreography.[13] In the textual arts, AIs now routinely write simple newspaper stories, such as sports reporting, that have a relatively standard format.[14] More ambitious literary applications are underway, such as the generation of novels or screenplays by AI systems.[15] And even where the results of the AI output are bizarre or non-sensical, human readers may impute to them profound or nuanced meanings, perhaps validating the concept of reader-response theory.[16]

Machine learning systems are finding equally broad application in fields of utilitarian innovation. They are increasingly used in the development of new technical innovations, such as circuit designs or the

[9] Sonia Katyal & Aniket Kasari, *Trademark Search, Artificial Intelligence, and the Role of the Private Sector*, 35 BERKELEY TECH. L.J. 501 (2020).

[10] JOANNA ZYLINSKA, AI ART: MACHINE VISIONS AND WARPED DREAMS 50–51 (2020) (describing the "Next Rembrandt" AI graphics project).

[11] *See, e.g.*, Siobhan Roberts, Tanya Basu, Charlotte Jee, & Patrick Howell O'Neill, *Machine Creativity Beats Some Modern Art*, MIT TECH. REV. (June 30, 2017), https://www.technologyreview.com/2017/06/30/150666/machine-creativity-beats-some-modern-art [https://perma.cc/NX52-2MYG].

[12] *See, e.g.*, Andrew R. Chow, *'There's a Wide-Open Horizon of Possibility.' Musicians Are Using AI to Create Otherwise Impossible New Songs*, TIME (Feb. 5, 2020), https://time.com/5774723/ai-music [https://perma.cc/65HH-GZUN].

[13] *See, e.g.*, Genevieve Curtis, *Dances with Robots, and Other Tales from the Outer Limits*, N.Y. TIMES (Nov. 5, 2020), https://www.nytimes.com/2020/11/05/arts/dance/dance-and-artificial-intelligence.html [https://perma.cc/R4W7-UUT5].

[14] *See, e.g.*, Stephen Beckett, *Robo-Journalism: How a Computer Describes a Sports Match*, BBC (Sept. 12, 2015), https://www.bbc.com/news/technology-34204052 [https://perma.cc/YJ6B-W6VV].

[15] *See, e.g.*, Richard Lea, *If a Novel Was Good, Would You Care If It Was Created by Artificial Intelligence?*, GUARDIAN (Jan. 27, 2020), https://www.theguardian.com/commentisfree/2020/jan/27/artificial-intelligence-computer-novels-fiction-write-books [https://perma.cc/S9SS-XGE8].

[16] *See generally* Peter J. Rabinowitz, *Reader-Response Theory and Criticism*, *in* THE JOHNS HOPKINS GUIDE TO LITERARY THEORY AND CRITICISM 606 (Michael Groden & Martin Kreiswirth eds. 1994) (explaining the construction of textual meaning by reader response).

design of mechanical devices.[17] In the chemical and biological sciences, AI systems identify drug targets, or develop molecular structures directed to particular drug targets.[18] They are deployed for sorting through large data sets to identify the most efficacious treatment regimes or to identify new applications and uses for known pharmaceutical products.[19]

AI systems are expected to have a substantial impact not only in the generation of new entities within the subject matter of intellectual property, but on the legal administration of rights arising out of such subject matter. Options are currently under active examination for AI-assisted or enabled examination of patent applications at the United States Patent Office.[20] The same is true for trademark applications.[21]

AI proposals have also begun to figure in the administration of intellectual property enforcement. For example, Libson and Parchomovsky have suggested that predictive analytics should be used in determining copyright infringement, matching the award of damages to the algorithmically predicted willingness to pay of a copyright defendant.[22] One could quickly extrapolate this proposal to other areas of intellectual property, particularly the calculation of actual damages or reasonable royalties in patent enforcement. Thus, every indication is that AI systems will likely become as ubiquitous in the development and administration of intellectual property as they are becoming across myriad other activities.

[17] *See, e.g.*, Sam Shead, *Google Claims It Is Using A.I. to Design Chips Faster Than Humans*, CNBC (June 10, 2021), https://www.cnbc.com/2021/06/10/google-is-using-ai-to-design-chip-floorplans-faster-than-humans.html [https://perma.cc/V5HL-SZLY].

[18] *See, e.g.*, Scott LaFee, *Artificial Intelligence Could Be New Blueprint for Precision Drug Discovery*, UC SAN DIEGO HEALTH (July 12, 2021), https://health.ucsd.edu/news/releases/Pages/2021-07-12-artificial-intelligence-could-be-new-blueprint-for-precision-drug-discovery.aspx [https://perma.cc/2V7G-TKFV].

[19] *See, e.g.*, Lauren Hinkel, *Deep-Learning Technique Predicts Clinical Treatment Outcomes*, MIT NEWS (Feb. 24, 2022), https://news.mit.edu/2022/deep-learning-technique-predicts-clinical-treatment-outcomes-0224 [https://perma.cc/32KN-5CY4].

[20] *See* Arti K. Rai, *Machine Learning at the Patent Office: Lessons for Patents and Administrative Law*, 104 IOWA L. REV. 2617 (2019).

[21] *See, e.g.*, Drew Hirshfeld, *Artificial Intelligence Tools at the USPTO*, USPTO (Mar. 18, 2021), https://www.uspto.gov/blog/director/entry/artificial-intelligence-tools-at-the [https://perma.cc/WK6Z-TE8Z].

[22] Adi Libson & Gideon Parchomovsky, *Toward the Personalization of Copyright Law*, 86 U. CHI. L. REV. 527 (2019).

II. RACE AND IP

At the same time that AI systems are gaining an increased purchase in the generation and administration of intellectual property, our understanding of intellectual property doctrine and practice is undergoing a substantial shift as we discover our own history of racial inequity. Although there has been a long tradition of scholarship examining justice and inequality within American intellectual property jurisprudence,[23] recent work has put new focus on and devoted new energy to examining the details of racial disparities in the field.[24] Some of this work extends previous critical doctrinal analyses, for example demonstrating unnoticed racial dimensions in patent law's requirement of non-obviousness,[25] or revealing the surprising prevalence of racially charged language in patent claiming and subject matter.[26]

In numerous areas of law, critical race inquiries have disclosed implicit doctrinal inequalities in the treatment of minority populations, demonstrating that these inequalities persist even after explicitly discriminatory policies were repealed.[27] Substantial work of a similar nature is now underway in intellectual property, demonstrating for example the mechanisms by which intellectual property has either disadvantaged or taken advantage of African American creators.[28] Recent scholarship has shown that, although ostensibly neutral on their face, copyright doctrines such as originality have in practice a disparate impact, valorizing appropriation from subordinated com-

[23] *See, e.g.,* Peter Lee, *Toward a Distributive Agenda for U.S. Patent Law*, 55 HOUSTON L. REV. 321 (2017); Margaret Chon, *Intellectual Property Equality*, 9 SEATTLE J. SOC. JUST. 259 (2010); Keith Aoki, *Distributive and Syncretic Motives in Intellectual Property Law (with Special Reference to Coercion, Agency, and Development)*, 40 U.C. DAVIS L. REV. 717 (2007).

[24] *See* Vats & Keller, *supra* note 3.

[25] Jonathan D. Kahn, *Race-ing Patents/Patenting Race an Emerging Political Geography of Intellectual Property in Biotechnology*, 92 IOWA L. REV. 353 (2007).

[26] Shubha Ghosh, *Race-Specific Patents, Commercialization, and Intellectual Property Policy*, 56 BUFF. L. REV. 409 (2008).

[27] *See generally* Angela P. Harris, *Racing Law: Legal Scholarship and the Critical Race Revolution*, 52 EQUITY & EXCELLENCE IN EDUC. 12 (2019) (surveying the trajectory of critical race concepts in legal scholarship).

[28] Robert Brauneis, *Copyright, Music, and Race: The Case of Mirror Cover Recordings* (July 22, 2020) (GWU Legal Studies Research Paper, Paper No. 2020-56), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3591113 [https://perma.cc/8W48-4GN7]; Kevin J. Greene, *'Copynorms,' Black Cultural Production, and the Debate Over African-American Reparations*, 25 CARDOZO ARTS & ENT. L.J. 1179 (2008); Kevin J. Greene, *Copyright, Culture & (and) Black Music: A Legacy of Unequal Protection*, 21 HASTINGS COMM. & ENT. L.J. 339 (1998).

munities while penalizing appropriation by subordinated communities.[29] Similarly, Anjali Vats has argued broadly that doctrinal conceptions of creation, infringement, ownership, and value in intellectual property law are deeply racialized.[30]

But much of the most recent analysis has come in the form of empirical investigation, either quantifying the deficit of minority participation in intellectual property systems, or qualitatively showing the historical trends and outcomes from past exclusionary practices. Recent work has examined the legacy of slavery in the American patent system,[31] as well as the history of racial violence, such as lynching, perpetrated against black inventors.[32] And although some scholars have speculated that copyright law might have contributed in some measure to the advancement of racial equality,[33] scholars of color who have examined the historical record have largely differed with this thesis.[34] Several studies of African American contributions to music and entertainment have suggested that copyright has been a source of inequitable treatment, rather than a remedy for it.[35] Similarly, several studies show the deployment of copyright and trademark doctrines over time fostering and reinforcing popular stereotypes caricaturing people of color.[36]

The historical disadvantage experienced by minority creators remains unremedied, as demonstrated by current metrics of participation in IP systems. Bell and co-authors have shown that innovation and patenting are closely tied to geographic location – which, given the historic connections in the United States between race, poverty,

[29] Betsy Rosenblatt, *Copyright's One-Way Racial Appropriation Ratchet*, 53 U.C. DAVIS L. REV. 591 (2019).

[30] ANJALI VATS, THE COLOR OF CREATORSHIP: INTELLECTUAL PROPERTY, RACE, AND THE MAKING OF AMERICANS (2020).

[31] Kara Swanson, *Race and Selective Legal Memory: Reflections on "Invention of a Slave,"* 120 COLUM. L. REV. 1077 (2020); Brian Frye, *Invention of a* Slave, 68 SYRACUSE L. REV. 181, 194 (2018).

[32] Lisa D. Cook, *Violence and Economic Activity: Evidence from African-American Patents 1870-1940*, 19 J. ECON. GROWTH 221 (2014) (finding linkage between declining African-American patenting activity and racial violence).

[33] Justin Hughes & Robert P. Merges, *Copyright and Distributive Justice*, 92 NOTRE DAME L. REV. 513 (2017).

[34] *See* Vats & Keller, *supra* note 3.

[35] *See, e.g.*, K.J. Greene, *Intellectual Property at the Intersection of Race and Gender: Lady Sings the Blues*, 16 AM. U. J. GENDER, SOC. POL'Y & L. 365 (2008); Olufunmilayo Arewa, *Blues Lives: Promise and Perils of Musical Copyright*, 27 CARDOZO ARTS & ENT. L.J. 547 (2010).

[36] Kevin J. Greene, *Trademark Law and Racial Subordination: From Marketing of Stereotypes to Norms of Authorship*, 58 SYRACUSE L. REV. 431 (2008).

and housing, translates to racial disparities.[37] Recent studies have also provided data confirming and quantifying the paucity of racial minority participation in proceedings before the US Patent Office,[38] the US Trademark Office,[39] and the US Copyright Office.[40] African American inventors are, for example, not only underrepresented in the percentage of patent applications filed before the USPTO,[41] they are statistically more likely to have the applications that they do file denied.[42] Strikingly, early studies indicate that minority patent applicants with non-racially associated names are no more likely to obtain a patent than applicants with racially associated names,[43] leaving open the possibility that either the requirements for patentability are skewed against the circumstances of minority inventors, or that implicit bias from less apparent racially associated determinants in the patent application is contributing to the significant denial of applications by minority inventors.

Comparable patterns of disparity emerge from data on the administration of copyright and trademark registrations. White authors are overall overrepresented in copyright registrations compared to population composition.[44] Hispanic authors on the other hand are significantly underrepresented.[45] The picture for African American authors is mixed, as they are most likely to register musical works but significantly less likely to register software or textual works.[46] And, while copyright registration is not required to obtain a copyright, in

[37] Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova & John Van Reenan, *Who Becomes an Inventor in America? The Importance of Exposure to Innovation*, 134 Q.J. ECON. 647 (2019).

[38] W. Michael Schuster, R. Evan Davis, Kourtenay Schley, & Julie Ravenschraft, *An Empirical Study of Patent Grant Rates as a Function of Race and Gender*, 57 AM. BUS. L.J. 281 (2021); Holly Fechner & Matthew S. Shapanka, *Closing Diversity Gaps in Innovation: Gender, Race, and Income Disparities in Patenting and Commercialization of Inventions*, 19 TECH. & INNOVATION 727, 729 (2018).

[39] *See* Miriam Marcowitz-Bitton, Deborah R. Gerhardt, & W. Michael Schuster, *An Empirical Study of Gender and Race in Trademark Prosecution*, 94 SO. CAL. L. REV. 1407 (2021) (showing data suggesting that Black and Latino minorities, but not Asian minorities, are underrepresented in USPTO trademark prosecution).

[40] Robert Brauneis & Dotan Oliar, *An Empirical Study of the Race, Ethnicity, Gender, and Age of Copyright Registrants*, 86 GEO. WASH. L. REV. 46 (2018).

[41] Schuster et al., *supra* note 38, at 287–88.

[42] *Id.* at 306.

[43] *See id.* at 282.

[44] Brauneis & Oliar, *supra* note 40, at 59.

[45] *Id.* at 60.

[46] *Id*. at 62–63.

the United States it is required to enforce a copyright,[47] so that the underrepresentation of creators of color in copyright registrations suggests that racial minorities are not deriving the full benefit of the copyright system. Similarly, although rights in trademarks arise at common law, there are significant advantages to federal registration, and registrations by Black and Latino trademark owners lags their expected percentage from the general population.[48] This underrepresentation of USPTO registrations likely places them at a disadvantage in accruing the benefits of the trademark system.

Such scholarship on intellectual property and racism is in a relatively nascent stage, but already has disclosed multiple indicators as to distortions ensconced in existing practice. We should expect that the path intellectual property has followed has been distorted by such biases; intellectual property criteria and doctrines that have grown up in the absence of participation by subordinated groups are unlikely to have incorporated the insights, experiences, and viewpoints of those groups. But historically biased intellectual property jurisprudence will not only be marred by the substantive exclusion of underrepresented groups—racial bias or exclusion leaves not merely lacunae where missing participants might have been found, but altered practices in their place. We have preliminary evidence of such distortions along dimensions of gender; copyright long excluded the fiber arts and other expressive "crafts" associated with female social roles;[49] patent law similarly takes little account of "feminine" ways of thinking and knowing.[50] The growing evidence of exclusion of other subordinated populations presages similar findings with regard to race or ethnicity; intellectual property concepts such as "nonobviousness" or "creativity" might look quite different today if they had been shaped within an epistemically diverse context.

[47] 17 U.S.C. §114(a) (2012); *see also* Fourth Estate Public Benefit Corp. v. Wall-Street.com, 139 S. Ct. 881, 892 (2019) (holding that copyright enforcement cannot commence until after registration issues).

[48] *See* Marcowitz-Bitton et al., *supra* note 39.

[49] Shelley Wright, *A Feminist Exploration of the Legal Protection of Art*, 7 CANADIAN J. WOMEN & L. 59 (1994).

[50] Dan L. Burk, *Do Patents Have Gender?*, 19 AM. U.J. GENDER SOC. POL'Y & L. 881 (2011); Dan L. Burk, *Feminism and Dualism in Intellectual Property*, 15 AM. U.J. GENDER SOC. POL'Y & L. 183 (2007).

III. BIAS IN AI APPLICATIONS

Given these trends in current intellectual property jurisprudence—the contemplation of future AI involvement, and the realization of existing racial disparities—some commentators have begun to consider the convergence of the two, in essence asking about foreseeable outcomes from this confluence of past and future practices.[51] Each of these trends in isolation would be worthy of rigorous examination; taken together they clearly merit even closer critical scrutiny. Initial forays into the intersection of these trends raise sensible concerns that proceed from what is currently known about the characteristics of each trend. Critical race theorists have already observed that "AI" as a cultural phenomenon is coded white.[52] The problem of bias in AI systems has already been the subject of considerable scrutiny, and numerous commentators have suggested that AI bias raises novel issues beyond our already sordid history of widespread social inequity.[53] In the context of intellectual property systems that are increasingly recognized as inherently discriminatory, some apprehension over the introduction of AI systems is therefore natural and sensible.

A. DISAMBIGUATING AI BIAS

But in order to consider the problem of AI bias in intellectual property systems—or for that matter, in any context—we must first clear up a degree of confusion in the current literature. Specifically, it is necessary to separate out different and sometimes confusing uses of the term "bias." Increasingly large swaths of the technical literature and of legal commentary concern the presence of detrimental "bias" in AI systems, but the use of the term is not consistent, and in many cases a careful reading reveals that commentators are discussing different phenomena.

One common use of the term bias in the context of machine learning refers to bias in a technical sense, having to do with improperly calibrated operational features of an AI system. For example, bias may

[51] *See, e.g.*, Robinson, *supra* note 4.

[52] Stephen Cave & Kanta Dihal, *The Whiteness of AI*, 33 PHIL. & TECH. 685 (2020).

[53] *See, e.g.*, Sonia Katyal & Jessica Y. Jung, *The Gender Panopticon: Artificial Intelligence and Design Justice*, 68 UCLA L. REV. 692 (2021) (discussing gender and sexual orientation biases in AI); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, (2020) (discussing AI bias in employment discrimination); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact,* 104 CAL. L. REV. 671 (2016) (surveying AI bias and discrimination law).

refer to statistical biases, or skew in the data set on which the algorithm operates.[54] The data used may encompass a non-representative sample or may otherwise either improperly include or exclude data points in ways that will distort the outcome of the analysis. This is a particular problem with training data that in effect sets the parameters for further analysis, as any subsequent analysis will be tainted, but may also occur in the selection or formatting of data chosen for analysis. Other types of design defects might also contribute to inaccuracy, such as an improper choice of statistical model, or incorrect selection of algorithmic operations that process the system's input.[55]

Now, to some extent we must recognize that all data is "biased" in the sense that all data must be acquired and manipulated in order to be put to any use. As Geoff Bowker famously observed, raw data is an oxymoron—data does not exist as an independent entity in the universe.[56] Although we may talk of "gathering" or "harvesting" data, it is in fact manufactured by deliberate selection, codification, curation, cleaning, and formatting.[57] The resulting data set inevitably carries the marks not only of the underlying phenomenon it is meant to describe, but of the choices made in processing the data to a useable form. This creates a particular danger in mixing or re-purposing data sets generated for one objective to be subsequently applied to a different objective—a practice that routinely occurs in AI training and analysis.

But typically, where bias is discussed in the technical sense, the question is not merely whether the data has been adapted to a particular use, but rather is a question of *inappropriate* technical bias—that is to say, not whether some choices were made in data selection or curation, but whether the choices rendered the data not "fit for purpose." Assuming all machine learning designs will be biased in some way, the question is then whether the biases are those that facilitate the purpose to which the system will be applied, or whether they are contrary to the application. As a general rule, such discussions focus

[54] Ramya Srinivasan & Kanji Uchino, *Biases in Generative Art–A Causal Look from the Lens of Art History*, PROC. 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 41, 47–50 (2021).

[55] *See, e.g.*, Laura Pedraza-Farina & Ryan Whelan, *A Network Theory of Patentability*, 86 U. CHI. L. REV. 63, 141–42 (2020) (couching bias concerns in terms of "biased data" or "biased measure-design").

[56] GEOFFREY C. BOWKER, MEMORY PRACTICES IN THE SCIENCES 184 (2005) ("Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care.").

[57] Julie Cohen, *The Biopolitical Public Domain: The Legal Construction of the Surveillance Economy*, 31 PHIL. & TECH. 213, 224–25 (2016).

on bias as a problem of AI *accuracy*, concerned that bias in data or design could lead to inaccurate outputs. Certainly, all data will be biased in a broad sense, and we in fact *want* the data to be biased in ways that makes it amenable to analysis. Thus, the implied or explicit solution for "bias" in this sense is to get better data or to improve the technical design—if the results obtained from AI systems are faulty, then of course we need to repair or improve the AIs.

To take one widely discussed example, numerous commentators have criticized the "AI bias" seen in the repeated failures by various AI-driven facial recognition systems to properly identify subjects with darker skin tones.[58] These errors have been traced to the use of training data drawn largely from Caucasian facial portraits, or to other software or hardware design flaws that assumed white subjects would be the "normal" or default population for recognition.[59] Consequently, the systems relying on such expectations failed in their intended purpose when confronted with a more diverse array of facial features falling outside their baseline assumptions. A natural response to such faults, intended to improve system accuracy, would be to use a more diverse set of training data portraits, or to make other software and hardware adjustments to allow compatibility with the features of darker skinned subjects.

In this example, the technology was not "fit for purpose" —racial assumptions made by the system designers and trainers resulted in faulty operational outcomes. Note that the operation of the system was grounded in ostensibly neutral and objective measurements of physical phenomena—the optical perception and spatial distribution of human facial features. But because the design of the system incorporated, probably unconsciously or implicitly, racial biases regarding the target population for the device, that design then implemented a racially disparate outcome: failure to identify and authenticate persons with unexpectedly darker skin tones.[60]

[58] *See, e.g.*, Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 7 (2018); David Leslie, Understanding Bias in Facial Recognition Technologies (Oct. 5, 2020) (arXiv:2010.07023), https://arxiv.org/abs/2010.07023 [https://perma.cc/P933-RURW].

[59] *See* Leslie, *supra* note 58, at 12–15.

[60] *See* WENDY HUI KYONG CHUN, DISCRIMINATING DATA: CORRELATION, NEIGHBORHOODS, AND THE NEW POLITICS OF RECOGNITION 185–96 (2021) (tracing in detail the history of racial bias, grounded in eugenics, that underlies current facial recognition systems).

The example therefore illustrates an additional dimension or corollary to Bowker's observation regarding the nonexistence of "raw" or unprocessed data, which is that all human artifacts incorporate and embody the values of their creators. This is not a novel or revolutionary observation, forming as it does the bedrock assumptions of fields such as archaeology or the humanities—that we can observe a sculpture, tool, or implement from another place or era and understand from its characteristics something of the thoughts, culture, and practices of those who created and used it. The same is true for more recent human implements; Ziploc storage bags, clothes hangers, and sport-utility vehicles all bear mute witness to the customs and values of our current civilization. And such qualities are as true of data as any other human artifact, meaning that data is quintessentially a human artifact.

B. AI AND SOCIAL BIASES

This observation brings us to a different set of concerns, also denominated as "AI bias," which are found in discussions of the ethical and social dimensions of AI deployment. If artifacts reflect the values of their creators, then AI systems can be expected to harbor such embedded social values, some of which will inevitably be unpalatable or negative. The corpus of past activities from which we might train AIs to assess future metrics were developed in an environment marred by prejudice, and we should expect that they will incorporate and reflect such biases. Similarly, the corpus of past creative works from which we might train AIs to generate future works were developed in an environment marred by prejudice and will similarly carry the marks of their origins. The notions of value, popularity, and merit by which we judge creative and innovative works certainly incorporate similar prejudices.

The implications of such value biases clearly constitute a set of concerns distinct from technical biases. Rather than being worried about technical *inaccuracy*, commentators addressing these separate issues tend to be concerned with social *inequity*—discriminatory effects or outcomes from algorithmic processes.[61] Thus, in the example above regarding defective facial recognition systems, we noted that technical biases resulted in technical error, but both the *source* of the

[61] *See* Sandra Wachter, Brent Mittelstadt, & Chris Russell, *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, 123 W. VA. L. REV. 735 (2021) (distinguishing technical bias from social bias).

technical biases and the *impact* of the technical error implicate something more than inaccuracy: the algorithmic implementation of social rather than technical bias.

Turning to the subject matter of intellectual property, and taking just one example among many, Hassine and Neeman have deplored the production of AI-generated "zombie art" from systems trained on the works of historical masters.[62] In the famous case of the "Next Rembrandt," where a novel painting in the style of Rembrandt was generated by an AI trained on Rembrandt's digitized paintings, they point out that the result displays none of Rembrandt's surprisingly progressive treatment of women and other subjects.[63] They argue that the AI's biased output resulted not only from the assumptions embedded in Rembrandt's work selected for training, but from choices made by the team that designed and implemented the project.[64] We may expect that similar embedded biases will manifest themselves in the automated generation of novel music, texts, or other expressive works as AIs develop future products from the data patterns of past successes.

Similarly, the fundamental principles and doctrines by which we administer intellectual property rights, either via human or AI administration, are likely to incorporate discriminatory elements in unexpected and sometimes undetected ways. For example, the criteria for patentability—novelty, utility, and nonobviousness—have been shaped in an environment that we are coming to realize was infected with racial biases.[65] I have previously argued that such criteria display the marks of sexism; we should expect that under careful examination they will prove to be equally racist.[66] AIs trained on past racist or sexist outcomes will inevitably reproduce such outcomes, some of which may be immediately apparent, but many of which may not be – recall that we are only now beginning to admit and understand aspects of the social biases long embedded in our intellectual property systems.[67]

Note that in the example discussed above, of facial recognition software maladapted to certain racial or ethnic groups, the technical

[62] Tsila Hassine & Ziv Neeman, *The Zombification of Art History: How AI Resurrects Dead Masters, and Perpetuates Historical Biases*, 11 J. SCI. & TECH. ARTS 28 (2019).

[63] *Id.* at 31.

[64] *Id.*

[65] *See* Ghosh, *supra* note 26, at 493–94.

[66] Dan L. Burk, *Diversity Levers*, 23 DUKE J. GENDER L. & POL'Y 25 (2015); *see also* Kahn, *supra* note 25 (examining racist assumptions in patent non-obviousness doctrine).

[67] *See* discussion *supra* Part II.

bias is closely intertwined with socially constructed facts and assumptions—not merely the unstated assumption that "standard" or "normal" facial features analyzed by the system would be Caucasian, but *also* the assumption that facial features are a sufficiently stable and consistent set of metrics to constitute a viable means of authenticating identity. In one sense, the technical biases of the system might be addressed by getting better and more diverse training data—the inaccurate operation of the system could be "corrected" to address the operational failure in recognizing darker-skinned individuals. But this correction still assumes that facial recognition is a feasible and desirable method of identifying *any* individual, and that is a social assumption that, if incorrect or problematic, cannot be corrected with better data.

Similarly, the biases manifest in the example of the "Next Rembrandt" painting are to some degree technical biases—a biased output generated from skewed or non-representative samples. But such technical bias reflects deeper social assumptions underlying the selection of the data and the selection of the project itself. Such assumptions include the choice of the digitized training data, which was drawn entirely from Rembrandt's paintings of white men (Rembrandt's paintings are not devoid of women or persons of color that might have been used).[68] The project is laden with multiple other assumptions, such as the premise that Rembrandt is an artist whose work should be emulated, and, of course, the premise that algorithmically generating paintings in the style of deceased artists is a laudatory or worthwhile application of machine learning technology.

Thus, to some extent the two sets of concerns may overlap. Skewed or biased statistical sampling may well lead to socially objectionable outcomes. But statistically sound samples may equally well lead to socially objectionable outcomes, especially if the underlying trends or practices are socially biased but not statistically or technically biased. For that matter, it is not impossible that technically biased AI systems might accidentally result in socially equitable results, depending on the nature of the particular bias in the system. The two sets of problems are not unrelated, but they are distinct, and solving one problem will not necessarily solve the other—indeed, as difficult as the problems involved in solving technical inaccuracy may be, they pale in comparison to the difficulties, possibly intractable difficulties, involved in solving problems of social inequity.

[68] *See* Hassine & Neeman, *supra* note 62.

C. Dispelling the Accuracy Fallacy

We have established that technical and social biases in AI pose two separate sets of questions, even if they use the same term—bias—to describe the respective problems. Nonetheless, these separate sets of concerns have tended to become muddled with one another in commentary oriented primarily toward one or the other. Critiques of social bias in algorithmic outcomes have in some cases seized on the *technical* literature dealing with bias to argue that algorithmic outcomes can or will suffer from *social* bias.[69] At the same time, technical literature addressing *statistical* or *design* biases often regards the resolution of such biases as solving the problem of *social* bias in AI systems.[70] The use of a common term contributes to mistakenly equating the two problems.

Some of the confusion in the algorithmic discussion surrounding social and technical bias stems from the conception of "accuracy" in different applications of machine learning. Accurate or inaccurate AI outputs are routinely, and mistakenly, equated with biased or unbiased AI outputs. Here it is important for us to distinguish between *social facts* and what I will call *natural* or *universal facts*. We accept that certain facts exist independent of whether humans observe them or are even aware of them—the speed of light in a vacuum, or the acceleration of a mass in earth's gravitational field will be constants no matter what humans think of them. There is perhaps a degree of social entanglement between these facts and human society, such as the units of measurement chosen, or even the choice to measure the phenomena at all. Some people may not accept such natural facts, or may not like them, but their existence, influence, and qualities are independent of human consideration or approval.

This is not the case with other sorts of facts, which have been called social facts, and which arise purely out of human association and society.[71] The fact that a piece of paper in my wallet, bearing an engraved portrait of Andrew Jackson, is worth twenty dollars is entirely dependent on human consideration and agreement.[72] There is

[69] *See, e.g.*, Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 59 (2019) (characterizing algorithmic bias in terms of reliability versus error).

[70] *See, e.g.*, Srinivasan & Uchino, *supra* note 54.

[71] Emile Durkheim, *What Is a Social Fact?, in* THE RULES OF SOCIOLOGICAL METHOD AND SELECTED TEXTS ON SOCIOLOGY AND ITS METHOD 50, 59 (Steven Lukes ed., W.D. Halls trans., 1982).

[72] JOHN R. SEARLE, THE CONSTRUCTION OF SOCIAL REALITY 189–90 (1995); Durkheim, *supra* note 71, at 59.

no universal phenomenon outside of human society called a "dollar." The concept of the "dollar," the concept that it has value for exchange, and the concept that it is represented by a certain stylized piece of paper are all the product of human social constructs. Dollars and other social facts exist because of human agreement, and they can be changed by human agreement, or abolished entirely if humans decide not to accept them.

This distinction is crucial in distinguishing between different types of AI assessments or predictions: there is a fundamental error in considering AI assessments of natural facts in the same way as considering AI assessments of social facts. From a technical standpoint the tasks in each case may appear much the same; each constitutes statistical modeling of correlations in large data sets. Data formatting, handling, and analysis; statistical modeling; and system design might appear much the same for either. It may seem quite straightforward to apply AI tools to each. But there is a substantial leap from using AIs for analysis of natural facts to using AIs for analysis of social facts. In particular there is a substantial gap with regard to the idea of "accurate" outcomes.

Using machine learning systems to trawl through large data sets in search of, say, astronomical evidence of black holes, or the presence of cancer cells in CT scans, is an exercise in searching for data profile of natural facts that have some independent existence. We expect that black holes, cancer cells, or other natural facts have a particular signature which, in relation to their characteristics, is either accurately perceived or not. The same is decidedly *not* true for assessments of social facts such as "substantial similarity" or "secondary meaning" or "patentable non-obviousness." These are entirely socially constructed facts like the concept of the "dollar." There is no question of measuring them accurately in regard to some independent baseline, because they are simply the product of social agreement and convention, and much like Gertrude Stein's Oakland, there is no there there.[73]

It may of course be that an AI system will be better or worse at assessing social constructs, in the sense of being more or less "accurate," at divining the occurrence of past constructions of social facts, based on the data that is available about that past practice. Depending on the design of the system, the data available, and the cooking of that data, a system may be more or less accurate in identifying what we have treated as "non-obvious" or "original" or "likely to confuse the ordinary observer" as those treatments are manifested in the records

[73] GERTRUDE STEIN, EVERYBODY'S AUTOBIOGRAPHY 289 (1937).

of such decisions. But, unlike Plank's constant or terminal velocity, such social facts have no independent valence; they are entirely malleable, change over time, and need not be the same in the future as they have in the past. What is being measured in such cases is only the implicitly or explicitly agreed-upon meaning of a social practice, not a stable and durable quantity.

A common response to issues raised regarding AI bias is some version of the argument that AIs will be more "accurate" than humans at making certain assessments, or that AIs will improve over time as their designs are perfected.[74] These responses assume that there is some natural or universal baseline against which the output of the algorithm can be measured. This conflates the different types of bias I have identified above—statistical, technical, and social[75]—and assumes that the problem in biased AI assessments is that of inaccurate outcomes, so that what is needed is to address the problem is better data or better analytical design. But even if AI systems could be made entirely free from statistical and inappropriate technical biases, or more accurate than human decision making, this does not solve or even seriously engage the social bias question. The divergence between these concerns should be clear—the argument from technical inaccuracy is somewhat beside the point if the ultimate concern is social bias in the sense of racial discrimination or disparate impact on subordinated social groups.

Thus, to bring this back to an IP example, the socially agreed-upon characteristic of patent "non-obviousness" is not a natural characteristic of some devices or technologies. It rather constitutes a policy lever that we have developed to overcome disincentives to technological development.[76] In the past we have typically defined it in a way that addresses epistemic or financial barriers to innovation.[77] But as I have argued elsewhere, there is no reason that it cannot be directed toward overcoming racial or gender classifications if those are posing

[74] *See, e.g.*, Alex P. Miller, *Want Less Biased Decision? Use Algorithms*, HARV. BUS. REV. (July 26, 2018), https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms [https://perma.cc/QRS4-UY5H] (asserting that algorithms "are less biased and more accurate than the humans they are replacing"); Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 570 (2018) (asserting that algorithms can be designed to avoid biases).

[75] *See* discussion *supra* Part III.

[76] Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 79 VA. L. REV. 1575 (2003).

[77] *See* Robert P. Merges, *Uncertainty and the Standard of Patentability*, 7 HIGH TECH. L.J. 1 (1992).

barriers to the "progress of the useful arts."[78] In particular, if the past practices with regard to "non-obviousness" (or other patenting requirements) has been constructed in such as way so as to systematically exclude from the patent system certain groups of innovators or their distinctive style of innovation, then the characteristics of the past practice, no matter how accurately divined, cannot be used as the future construction of such terms if we wish the future to be more diverse and inclusive.

IV. ADDRESSING AI SOCIAL BIASES

Drawing on the larger pre-existent literature on AI biases, prior commentary identifying the problem of AI bias in IP has recommended a suite of curative measures, suggesting for example that algorithmic transparency or human oversight of AI decisions is needed to ameliorate racial biases that may creep into applications of AI to intellectual property administration.[79] These are the solutions routinely proposed to ameliorate concerns about AI bias in other settings. But such proposed remediations seem not merely inadequate to address the problems raised, but appear largely directed at the wrong set of problems altogether. They are, again, largely directed to the problem of algorithmic accuracy. If the process is transparent, then mistakes can be identified, and accuracy improved. If a human is reviewing the AI outputs, then anomalies that might go unrecognized by machine "intelligence" can be identified by human intelligence for investigation and correction.

Consider for example the recommendation of designing systems with a "human in the loop"; that is, incorporating a layer of review and approval of AI outputs by a human overseer.[80] Rather than automating the entire system, leaving the machine quite literally to its own devices, a human would be charged with examining the machine's output to identify and correct biased recommendations before they were implemented. But careful consideration of this approach reveals a number of potential defects as a solution for algorithmically perpetuated social bias. First, as a practical matter, the economic forces propelling us toward deployment of AIs disfavor the incorporation of "humans in the loop." Much of the appeal of AI is the ability to *remove* humans from the loop, replacing relatively expensive and slow human

[78] Burk, *supra* note 66.

[79] *See* Robinson, *supra* note 4, at 764, 768–69.

[80] *Id.* at 768–69.

work with faster and cheaper automated work.[81] Placing human review back into the process largely negates the advantages that make AI engagement attractive. Indeed, given that humans currently constitute the entire loop, it is unclear in many cases why we would bother with engaging AIs if the process requires human oversight—we already have plenty of humans in the loop, with whatever advantages or disadvantages that entails.[82]

There may of course be situations where partial automation would be attractive, using AIs to advise or to supplement human direction. But even in these cases, relying on human oversight to correct bias assumes that the human will recognize the algorithmic bias and will not introduce additional bias into the loop. Given that the current biases in the system result from human prejudice, witting or unwitting, neither is it clear that putting a human in the loop would solve the algorithmic bias problem—the algorithmic bias problem arises from data generated by the humans who currently *are* the loop.[83] The examiner who is drafted to watch for algorithmic bias is presumably the same examiner who currently, for whatever reason, is less likely to approve the patent application of a female or minority inventor.[84] Assigning an examiner to oversee the algorithmic decision might prove corrective if we had unbiased examiners to do the job, but if we had unbiased examiners available, algorithmic bias would either be less of a problem, or not a problem at all. Thus the "human in the loop" solution seemingly re-introduces all the mistakes and delays of the current arrangement.

To illustrate the distinction, I adopt again a well-documented example from outside of intellectual property, the use of actuarial systems to predict recidivism for parole determinations.[85] One of the key factors that is used in predicting criminal recidivism in such systems

[81] *See, e.g.*, Bernard Marr, *The Economics of Artificial Intelligence—How Cheaper Predictions Will Change the World*, FORBES (July 10, 2018), https://www.forbes.com/sites/bernardmarr/2018/07/10/the-economics-of-artificial-intelligence-how-cheaper-predictions-will-change-the-world/?sh=55fcc5985a0d [https://perma.cc/R5EJ-JQMM].

[82] *Cf.* Dan L. Burk, *Algorithmic Fair Use,* 86 U. CHI. L. REV. 283, 300–01 (2019) (making this point in the context of automated fair use determinations).

[83] *See* discussion *supra* Part III.B.

[84] *See* Schuster et al., *supra* note 38.

[85] *See* Jessica Eaglin, *Technologically Distorted Conceptions of Punishment*, 97 WASH. U. L. REV. 483 (2018).

is the subject's postal zip code.[86] Dwelling location is considered highly predictive of criminal arrest, arraignment, and conviction.[87] Sadly, it does not require the assistance of an AI to detect the correlative pattern that might emerge from zip code data: in the United States, housing location is closely aligned with poverty, poverty is closely correlated to race, and racial minorities—particularly young African American males—are disproportionately enmeshed in the criminal justice system. Connecting the dots among these factors leads to a depressingly predictable conclusion.[88] In effect, the algorithm engages in a form of redlining, relating criminal behavior to geographic location.

For our purposes here, the take-away message from such predictions should be that algorithmic detection of these correlations is almost certainly not the result of any technical bias, although it demonstrates profound social and ethical bias. The prediction is not wrong in the sense of being inaccurate—to the extent that an actuarial system detects a correlation between zip codes and criminal behavior, the machine is not mistaken. Quite to the contrary, the problem that should concern us is that the correlation *is* quite accurate. Our concern should not be whether the predictive system is well designed, or whether the data is skewed by selective sampling or some other statistical impropriety. We instead need to question why the correlation exists and why it is permitted to persist. Critiquing its accuracy is the wrong critique.

Stated differently, in terms of a common concern over predictive analytics, the argument from technical accuracy might cause us to fret over whether the algorithm can identify persons living within the zip code who might not be inclined to recidivism—over whether the correlation is imprecise such that it sweeps into its ambit particular cases for which the correlation does not hold. That is of course always a problem with actuarial predictions, that not every data point fits the curve. But while we should not be unconcerned about such errors,

[86] *See, e.g.*, Nancy Ritter, *Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise,* NIJ JOURNAL 4, 6 (February 2013) (discussing incorporation of zip codes into predictive analytics as among the strongest predictors for recidivism).

[87] *See, e.g.*, Richard Berk, *The Role of Race in Forecasts of Violent Crime*, 1 RACE SOC. PROB. 231, 237 (2009) (showing strong statistical correlation between zip code and recidivism); Charis E. Kubrin & Eric A. Stewart, *Predicting Who Reoffends: The Neglected Role of Neighborhood Context in Recidivism Studies*, 44 CRIMINOLOGY 165 (2006) (finding strong correlations between criminal recidivism and parole to "disadvantaged" neighborhoods).

[88] *See* Berk, *supra* note 87 at 232 ("In locales with substantial residential segregation, knowing the zip code is virtually the same as knowing an individual's race.").

they overlook to a substantial degree the glaring problem inherent in the prediction's accuracy. We need to ask as an initial matter why any point *does* fit the derived curve, and why we are using the actuarial prediction at all. Our concern in deploying such algorithms should be why zip code is a substantial predictor of criminal arrest in the first place, and why we tolerate the conditions that make it so.

If we are concerned about building systems that are increasingly accurate in identification of criminal recidivism, in our current social setting we are in effect worrying about building systems that are increasingly adept at racial profiling—our current situation is that the trifecta of race, poverty and crime are closely tied together. The predictive algorithm in this instance is attuned to factors, such as race, that have become embedded in the concepts it was deployed to assess. This is an ugly and unfortunate truth to confront, but geography and its racial correlatives are part of the definition that we have constructed of "recidivism" and "crime" in the United States.

Moving from existing literature on actuarial criminal justice to the emerging discussion on actuarial intellectual property, we should be able to discern a similar set of potential concerns. If, let us say, the AI patent examiner, or the AI-assisted human patent examiner, illegitimately identifies illegitimate characteristics in an application and relies on them to require rejection or claim narrowing, that should not surprise us—we know that such implicit bias is already present in the patent system.[89] Those past practices unquestionably inform any data we might use to train future AI examination tools. Neither are those analytics biased in the technical sense of being inaccurate. Like the correlation between zip code and crime, they may very well be entirely accurate in the sense of predicting what we have in fact come to expect for patentability—because our concept of patentability has come to include such bias.[90] We would reject such outcomes not because the AI has inaccurately assessed our practices, but rather because we dislike the image we see in the mirror AI holds up to our practices.

To be certain, there is an ongoing discussion and a literature on "debiasing" data for AI analysis, for example looking for gender stereotypes in texts and attempting to substitute or correct for gender neutral language.[91] Such attempts assume that the corrections and tweaks

[89] *See* discussion *supra* Part II.

[90] *See* Shuster et al., *supra* note 38, at 317–18.

[91] *See, e.g.*, Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama

needed to debias the data are discrete and discernable, which they may not be. They concomitantly assume that we know what an "unbiased" baseline against which to evaluate the existing data would look like. At their core, they naively assume that data and the processes within which it is gathered, evaluated, and analyzed can be somehow isolated from the broader sociotechnical networks in which they are situated. But such attempts to clean up the data do not (and indeed cannot) address the social system in which the system is embedded.

Where IP is concerned, we might, for example, try to identify clues in patent applications that indicate race or gender, and attempt to remove or correct such clues from consideration by an AI examination system. But if our concept of, say, "nonobviousness" is in part defined by past racial prejudice that has become inherent in the patentability standard, applications that diverge from that standard will be still excluded, just as they have been in the past, whether or not they include overt markers of disadvantage. The immanence of reiterated social bias is simply implicit in the project of predicting or recommending future action based on correlations to past social behaviors. The analytic product of data garnered from social activity that incorporates structural inequalities must inevitably itself bear the hallmarks of such inequality.

V. DISTINGUISHING AI BIAS

This analysis brings to the fore a suggestion made by Anupam Chander: that concerns about AI bias, or discrimination arising from machine learning, are to some extent superfluous and a distraction.[92] Chander suggests that vetting the origins of discriminatory practices is less important than correcting their discriminatory effects—the problem we face is improper biases and discrimination; the source is somewhat beside the point.[93] We currently have unacceptable racial biases in our intellectual property systems; implementation of AI systems threatens to perpetuate such bias. Whether the bias comes from human or machine, or machines mimicking the more unsavory aspects of human society, we can more easily recognize biased outcomes than biased inputs. In order to eliminate racial disparity, whether it

& Adam T Kalai, *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings, in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 4356 (2016).

[92] Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1039–41 (2017) (advocating algorithmic "affirmative action" to cure discriminatory AI outcomes).

[93] *Id.* at 1024–25.

comes from a human or from a machine trained on the previous actions of humans, the solution is to identify disparate outcomes and correct them, no matter where the disparity originates. Chander's approach would shift the focus to discriminatory or disparate outcomes in IP systems, rather than fretting over the origins of such outcomes.[94] We would correct biased AI results as we would correct those that are becoming apparent in the current system.[95]

This suggestion seems at first blush eminently sensible, particularly given the longstanding and intractable nature of biased and disparate outcomes in our social institutions. It sidesteps the sticky problems of AI opacity and implicit bias to focus on solutions rather than on diagnosis. But on closer consideration, this approach can be sensible only to the degree that we believe that human-implemented biases are the equivalent of machine-implemented biases. If one sort of biased outcome is identical to the other, then we can perhaps solve them both in the same way. But if to the contrary the biased product of AI analysis differs in some substantial respect from our usual human biases, then treating all biases the same simply will not do. In particular, if we have reason to suspect that machine-driven discrimination may be more virulent or persistent than human-driven discrimination, then we will need to draw distinctions between our responses to each. If bias originating in or perpetuated through AI systems is somehow different, or potentially more problematic, then it may call for different or more drastic solutions than those we might deploy against existing biases endemic to intellectual property.

I suggest that algorithmic bias does diverge from direct human bias in at least two aspects that make a difference in how we can and should deal with inappropriate outcomes from AI systems. The first of these is the illusion of objectivity that surrounds algorithmic systems.[96] Humans are well-versed in assessing the actions of their fellow humans, and are accustomed to scrutinizing such actions for bias, favoritism, or normative lapses. However, humans are far less adept at assessing the outcomes of algorithmic determinations, imputing to them a degree of objectivity and neutrality that they would typically

[94] *Id*. at 1039.

[95] *See, e.g.*, *id*. at 1043.

[96] Tarleton Gillespie, *The Relevance of Algorithms*, *in* MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY 167, 179 (Tarleton Gillespie, Pablo J. Boczkowski & Kirsten A. Foot eds., 2014).

not assign to similar determinations by a person.[97] Fostering this human tendency works in the interests of purveyors of such systems, who may frame algorithmic systems in terms of objectivity in order to advance their own agendas—perhaps simply to sell more machine learning systems, or perhaps to quell dissent against advantageous outcomes from such systems.[98]

But whether or not it is advanced with ulterior motives, such framing plays into a broader tendency to assume that AIs will be more objective than humans, or at least that they may be more objective than humans if only the machines are properly designed and deployed. The illusion of AI objectivity draws upon characterizations of AI systems in which human participation, manipulation, and operation is placed outside the frame of consideration, making the devices appear to be "autonomous" in a misleadingly strong sense of that term.[99] We blithely talk for example of deploying "autonomous vehicle systems," as if the vehicle will be somehow imbued with the gift of self-determination to function free of human input. In fact, what we mean is that it will be free of human control in the very limited sense that a human will not be directly steering the course of the vehicle. But the label of "autonomy" tends to obscure the legions of humans who will be designing, manufacturing, deploying, maintaining, repairing, fueling, and otherwise supervising such vehicles. Similarly, when the ubiquitous human interventions in other AI systems, such as predictive analytics, are likewise obscured, the output of such analytics appears deceptively free from human control or manipulation, leading to unwarranted deference to such outputs.

Additionally, illusory AI objectivity is also heightened when the output or input to such systems is numerical. Humans have a strong propensity toward unwarranted deference to numerical systems generally; numbers appear authoritative and objective and are accorded a degree of influence that the same output or conclusion expressed in another language—say, English—would not be granted.[100] Once

[97] Jaap J. Dijkstra, Wimm B.G. Liebrand & Ellen Timminga, *Persuasiveness of Expert Systems*, 17 BEHAV. & INFO. TECH. 155 (1988). Human judgment is a complex process; in general people appear to view their own decisions as superior to those of algorithms, but algorithmic decisions to those of *other humans*. Jennifer M. Logg, Julia A. Minson & Don A. Moore, *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment*, 151 ORG. BEH. & HUM. DECISION PROCESSES 90, 94 (2019).

[98] Gillespie, *supra* note 96.

[99] *See* Burk, *supra* note 5, at 318–19.

[100] *See* danah boyd & Kate Crawford*, Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO. COMM. & SOC'Y 662 (2012).

again, this attribution of authority occurs in large part due to the framing of numbers without human context.[101] Numerical expressions appear to stand alone, lacking the relational connections that in another linguistic expression might trigger scrutiny, skepticism, or distinction.[102] Numbers are consequently perceived as unbiased and immanently meaningful. Many accept, as might be colloquially said, that "the numbers speak for themselves."

Of course, numbers do no such thing; no numerical expression exists in a cultural or social vacuum.[103] To assume that they "speak for themselves" is simply to ignore the context in which they are selected, devised, and deployed.[104] Far from standing on their own, numerical expressions are deeply value-laden, just as we have observed the AI systems that generate numerical outputs must inevitably be.[105] But the tendency to treat each of these as objectively abstract covers a multitude of biases that might be immediately detected and challenged if coming from a direct human source. Consequently, we may expect that biases emanating from AI systems will be less likely to be challenged, more likely to be excused, and more likely to be accepted, than the types of discrimination we currently experience.

When contemplating the distinctions between human biases and AI biases, such ill-placed confidence in AI objectivity would by itself be cause for heightened concern. But it is accompanied by additional distinctions from the sorts of bias that humans and human institutions are accustomed to addressing. I suggest that a second and related distinction between the bias entailed in familiar human activity and that entailed in AI systems is the degree of *performativity* associated with the latter.[106] The type of actuarial systems entailed in AI technology

[101] Marion Fourcade & Kieran Healy, *Categories All the Way Down*, 42 HIST. SOC. RSCH. 286, 292–93 (2017).

[102] *See* THEODORE M. PORTER, TRUST IN NUMBERS: THE PURSUIT OF OBJECTIVITY IN SCIENCE AND PUBLIC LIFE (1995).

[103] *See* boyd & Crawford, *supra* note 100, at 667 ("Claims to [numerical] objectivity are necessarily made by subjects and are based on subjective observations and choices.").

[104] *See* Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1685–86 (2020) (debunking the AI trope that "numbers speak for themselves").

[105] *See* Elish & boyd, *supra* note 6, at 69–70 (describing the epistemic construction of AI data models).

[106] *See* Irene Rafenell, *Durkheim and the Performative Model: Reconfiguring Social Objectivity, in* SOCIOLOGICAL OBJECTS: THE RECONFIGURATION OF SOCIAL THEORY 59, 62–66 (Geoff Cooper, Andrew King & Ruth Rettie eds., 2009) (tracing the development of performativity theories).

have consistently been observed to be socially performative, which for our purposes here I will define as *creating their own social facts* and *enacting what they assume*.[107] In particular, we should be concerned that AI systems will enact whatever social biases are entailed in their operational design and analytical subject matter.[108]

In explaining how and why this concern arises, I return to the differing classes of use for machine learning systems when applied to the analysis of data encompassing different types of facts. Above we distinguished natural or universal facts from social facts, the former having independent existence apart from human contemplation, and the latter being the product of human social agreement.[109] Unlike determined natural facts, social facts are dynamic, malleable, and most importantly, *open to alteration by the very process of analysis*. Because social facts are constituted entirely from human agreement, and algorithmic analysis may well alter that agreement, such analysis may substantially change the construction of the facts under consideration. We do not expect cancer cells or black holes to alter their behavior, or otherwise react to contemplation in actuarial models. But human subjects and institutions decidedly do alter their character from such scrutiny, and in particular will tend to adopt whatever assumptions are built into the instrument of scrutiny.

Taking once again a well-studied example from outside of intellectual property law, we can illustrate this type of performativity in so-called "predictive policing," where AI analytics are used to determine the likely location of crimes so that police resources can be deployed to that location in advance.[110] Predictions are developed on the basis of past criminal activity, identifying "hot spots" where crime has previously been reported.[111] Such predictions are therefore highly dependent on the quality and nature of the reporting.[112] Perhaps not surprisingly, reports of crime go up for locations where there are numerous police, since there are police there to observe and report such

[107] Dan L. Burk, *Algorithmic Legal Metrics*, 96 NOTRE DAME L. REV. 1147, 1170 (2021).

[108] *See, e.g.*, Cave & Dihal, *supra* note 52 (arguing that the design of AI systems includes affordances that are culturally coded as "white").

[109] *See* discussion *supra* Part III.C.

[110] Lyria Bennett Moses & Janet Chan, *Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability*, 28 POLICING & SOC'Y 806, 813 (2018); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 113–14 (2017).

[111] Moses & Chan, *supra* note 110, at 807–08.

[112] Mareile Kaufmann, *Who Connects the Dots? Agents and Agency in Prediction Algorithms, in* TECHNOLOGY AND AGENCY IN INTERNATIONAL RELATIONS 141 (Marijn Hoijtink & Matthias Leese eds., 2019).

crimes.[113] When police are deployed to areas where crime is expected, they observe and report crimes in the area, fulfilling the prediction by the algorithm. Such reports additionally reinforce predictions of crime for that location by supplying crime data for the next iteration of analysis.[114]

But not only will the data sampling be skewed by following such algorithmic predictions, the character of the neighborhood in question is likely to change due to the actuarial characterization. The population will react to increased police surveillance, potentially raising tensions and creating resentment. Officers stationed in the neighborhood may police differently given that it is a "high crime" neighborhood. Property values are likely to fall given the characterization; mortgages, automobile insurance, and other financial transactions may be affected. Residents who dislike intensive police surveillance and can afford to move may do so. Impoverished occupants who cannot afford to move, and who are more likely to become enmeshed with the criminal justice system, will remain. Thus, the prediction of the algorithm that crime will be plentiful becomes actualized due to the algorithm's prediction.

We should expect the same cyclic effects when AI systems are applied to intellectual property. For example, if we wish to use AI to determine "non-obviousness" —an entirely socially fabricated concept if there ever was one—in patent applications, the source for training data will inevitably be the set of non-obviousness determinations from past patent applications. AI examiners or examination aids trained on past findings of "non-obviousness" will construct from the data set indicia of non-obviousness based on past practice. The characteristics of those patterns will then be the ones then sought to be identified in future applications. The patent applications found to be non-obvious based on those criteria will supply the data for further, additional iterations of AI analysis. As these same criteria define successive generations of data on "non-obviousness," the definitional criteria from past practice are selected for and reinforced.

More importantly, applicants seeking successful issue of patents will highlight those selected characteristics in order to successfully prosecute their applications to issue, further valorizing and emphasizing the algorithmically determined criteria. Innovations that fail algorithmically entrenched definitions of non-obviousness may be kept as

[113] Moses & Chan, *supra* note 110, at 810.

[114] Selbst, *supra* note 110, at 141.

trade secrets, dedicated to the public domain, or simply never developed at all—patents are after all intended as incentives to investment in "non-obvious" inventions, and whatever we determine that "non-obvious" means is what we should expect to get more of. Thus, the algorithm will not merely select for applications that are "non-obvious," it will effectively define what that term comes to mean.

With regard to the biases of concern in this Essay, innovation facilitated by AIs or examination of patents via AI similarly threatens to entrench biases we are now uncovering in the patent system.[115] Thus, if the AI-determined meaning of non-obviousness includes indicia inadvertently tied to characteristics such as race or gender, those will be selected for, and the meaning of "non-obvious" will shift to include those characteristics. Note, too, that in this case transparency of the algorithmic practice may actually feed the self-fulfilling algorithmic prophecy by revealing to patent applicants the characteristics they should emphasize and display in their applications in order to successfully comport with the non-obviousness requirement.

Performativity is of course neither unique to nor limited to automated actuarial systems. It has been observed and documented in a variety of settings prior to and apart from the deployment of AI.[116] Consequently, our concern should not be so much that AIs will introduce these effects into intellectual property systems; rather, our concern should be that deployment of AI will magnify the practices already at work in human institutions. Because of cheap, fast, and ubiquitous computing power, AI systems are being deployed widely, and as with all digital automation, these systems may be expected to amplify and accelerate the processes they engage—in this case, familiar social processes, including detrimental and counterproductive social processes. Just as AI tools extend human cognitive ability, parsing data sets to identify patterns beyond human perception, so too they extend existing social practices such as performativity, enhancing those practices for good or ill.

When combined with the illusion of objectivity, we can expect that the self-fulfilling actuarial prophecies of algorithmic intellectual property will be less likely to be questioned or disputed. And those unquestioned performative outcomes will inevitably entail racial biases. We know very well that human actors are frequently the victims

[115] *See* discussion *supra* Part II.

[116] *See* Rafenell, *supra* note 106; Nicolas Brisset, *The Future of Performativity*, 7 ECONOMIA 439, 443 (2017).

of their own unrecognized implicit biases;[117] when those biases are embedded in automated systems that have no capacity for self-reflection or social awareness, and to which human observers impute undeserved credence, AI bias will not be bias as usual. Algorithmic IP will instead compound and amplify the problematic trends and outcomes already identified in intellectual property systems.

CONCLUSION

I have argued that the deployment of AI systems in the creation and administration of intellectual property will inevitably carry with it the racial biases we have begun to identify in IP systems. Moreover, such biases are not a problem of algorithmic accuracy or inaccuracy that can be solved by implementing more accurate designs or procuring more accurate data. Neither is it likely that such biases will be cured by watchful adjustments of algorithmic outcomes. I forecast that such biases may be especially pernicious because of the pervasive attribution of neutrality to numerical and technical systems, and the performative nature of algorithmic assessments. On the contrary, as with other digital technologies, AI systems may be expected to amplify and accelerate the objectionable trends we are now identifying in current practice and doctrine.

This bodes poorly for the use of AI for substantive deployment in IP creation or administration. At the same time, the propensity for algorithmic systems to amplify and replicate existing biases might, under the right conditions of deployment, constitute a feature rather than a bug. The use of algorithmic metrics in the examples that I have offered above helped to reveal and disclose racially biased practices that might otherwise have gone unnoticed or unappreciated. Wendy Chun has therefore suggested that machine learning systems might usefully lend themselves to *diagnosis* of social bias, rather than as guidance or implementation for social systems.[118] Chun compares such diagnostic uses of algorithms to a social analog of weather forecasting, where we use sophisticated modeling to understand and predict weather patterns but are never hoodwinked into believing that the model we have constructed is in fact the weather, or determines the weather.[119]

[117] *See* Kristin A. Lane, Jerry Kang & Mahzarin R. Banaji, *Implicit Social Cognition and Law*, 3 ANN. REV. L. & SOC. SCI. 427 (2007).

[118] CHUN, *supra* note 60, at 122.

[119] *Id.* at 122–23.

Using AI in this fashion might allow us to diagnose and re-orient racially tainted IP doctrines and practices toward greater inclusion and equity. Limited deployment of AI as a tool to disclose social bias in IP in essence turns the malleability of socially constructed facts to advantage. For example, returning to my illustration of non-obviousness above, it might be useful to deploy AI diagnostics to identify markers of racial bias in the concept of non-obviousness—but this is quite a different matter than trying to use AIs to identify or define non-obviousness. This diagnostic approach would require re-thinking and reformulating the role of AIs in IP, from definitional to investigative, from determinant to corrective. But rather than allowing AIs to ensconce socially biased qualities in intellectual property for the future, we might re-deploy it to reveal the defects in intellectual property systems now.