NORTH CAROLINA JOURNAL OF LAW & TECHNOLOGY
VOLUME 19, ISSUE 1: OCTOBER 2017

TAMING THE GOLEM: CHALLENGES OF ETHICAL ALGORITHMIC DECISION-MAKING

Omer Tene & Jules Polonetsky*

The prospect of digital manipulation on major online platforms reached fever pitch in the last election cycle in the United States. Jonathan Zittrain's concern about "digital gerrymandering" found resonance in reports, which were resoundingly denied by Facebook, of the company's alleged editing of content to tone down conservative voices. At the start of the last election cycle, critics blasted Facebook for allegedly injecting editorial bias into an apparently neutral content generator: its "Trending Topics" feature. Immediately after the election when the extent of dissemination of "fake news" through social media became known, commentators chastised Facebook for not proactively policing user-generated content to block and remove untrustworthy information. Which one is it then? Should Facebook have employed policy-directed technologies or should its content algorithm have remained policy-neutral?

This article examines the potential for bias and discrimination in automated algorithmic decision-making. As a group of commentators recently asserted, "[t]he accountability mechanisms and legal standards that govern such decision processes have not kept pace with technology." Yet this article rejects an approach that depicts every algorithmic process as a "black box" that is inevitably plagued by bias and potential injustice. While recognizing that algorithms are man-made artifacts, written and edited by humans in order to code decision-making processes, the article argues that a

*Omer Tene is a Senior Fellow at the Future of Privacy Forum, VP of Research at the International Association of Privacy Professionals, and Affiliate Scholar at the Stanford Center for Internet and Society. Jules Polonetsky is CEO of the Future of Privacy Forum and Adjunct Professor at the Washington & Lee University School of Law. We thank Kelsey Finch for helpful research and comments.

distinction should be drawn between "policy-neutral algorithms," which lack an active editorial hand, and "policy-directed algorithms," which are intentionally framed to further a designer's policy agenda.

Policy-neutral algorithms could, in some cases, reflect existing societal biases and historical inequities. Companies, in turn, can choose to fix their results through active social engineering. For example, after facing controversy in light of an algorithmic determination to not offer same-day delivery in low-income neighborhoods, Amazon nevertheless recently decided to provide those services in order to pursue an agenda of equal opportunity. Recognizing that its decision-making process, which was based on logistical factors and expected demand, had the effect of facilitating prevailing social inequality, Amazon chose to level the playing field.

Policy-directed algorithms are purposely engineered to correct for apparent bias and discrimination or to advance a predefined policy agenda. In this case, it is essential that companies provide transparency about their active pursuits of editorial policies. For example, if a search engine decides to scrub results clean of opposing viewpoints, it should let users know they are seeing a manicured version of the world. If a service optimizes results for financial motives without alerting users, it risks violating FTC standards for disclosure. So too should service providers consider themselves obligated to prominently disclose important criteria that reflect an unexpected policy agenda. The transparency called for is not one based on revealing source code but rather public accountability about the editorial nature of the algorithm.

The article addresses questions surrounding the boundaries of responsibility for algorithmic fairness and analyzes a series of case studies under the proposed framework.

INTRODUCTION

In the spring of 2016, six months before the United States presidential election, public outrage broke around reports that Facebook allegedly edited its "Trending Topics" feature to suppress conservative views.[1] Closer examination of the criticism leveled at

[1] Danah Boyd, *Facebook Must Be Accountable to the Public*, DATA & SOC'Y POINTS (May 13, 2016), http://bit.ly/1Xw14dm; Kashmir Hill, *Maybe the Real Facebook Suppression Is of Shoddy News, Not Conservative News*, FUSION (May 11, 2016, 6:40 PM), http://fusion.net/story/301156/facebook-suppression-

Facebook and other companies in various contexts discussed below demonstrates the complexity of our expectations for algorithms—computer programs written as sets of step-by-step instructions—which increasingly determine what information we are exposed to and what decisions are made about us.[2] Websites and social media platforms commonly provide curated lists of content, and therefore some users assumed that a mechanized algorithm automatically populated the box of "Trending Topics" shown to the right of their News Feed. Even a report on the tech news site Gizmodo asserted that Facebook editors had intentionally suppressed news topics from conservative publications trending across the network and inflated the importance of other favored topics by injecting them into user view.[3] This came on the heels of another report that Facebook employees asked CEO Mark Zuckerberg if the company had a responsibility to "help prevent President Trump in 2017."[4]

A public uproar ensued. For some users, the suspicion of Facebook editors advancing a political viewpoint was hard to swallow. Senate Republicans sent an angry letter to Facebook, requesting clarifications from Facebook as to whether there was any level of subjectivity associated with the Trending Topics section.[5] Facebook denied the reports several times, including in a post by Zuckerberg himself, who stated, "[w]e have rigorous guidelines that

---

conservative-news-shoddy-news/; Michael Nunez, *Former Facebook Workers: We Routinely Suppressed Conservative News*, GIZMODO (May 9, 2016, 9:10 AM), http://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006.

[2] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PENN. L. REV. 633 (2017).

[3] Farhad Manjoo, *Facebook's Bias Is Built-in, and Bears Watching*, N.Y. TIMES (May 11, 2016), http://nyti.ms/1ZVqsrx.

[4] Michael Nunez, *Facebook Employees Asked Mark Zuckerberg If They Should Try to Stop a Donald Trump Presidency*, GIZMODO (Apr. 15, 2016, 2:40 PM), http://gizmodo.com/facebook-employees-asked-mark-zuckerberg-if-they-should-1771012990.

[5] *See generally* Michael Nunez, *Senate GOP Launches Inquiry into Facebook's News Curation*, GIZMODO (May 10, 2016, 12:34 PM), http://gizmodo.com/senate-gop-launches-inquiry-into-facebook-s-news-curati-1775767018.

do not permit the prioritization of one viewpoint over another or the suppression of political perspectives."[6] In an unprecedented move, Facebook published its internal editorial guidelines, a 28-page document that details how editors and algorithms interact in the process of selecting "Trending Topics" on the website's feed.[7] Facebook's global policy chief, Joel Kaplan, himself a prominent conservative, blogged about Facebook's role as a platform enabling conservative voices to spread their message, and reports from social media tracking companies confirmed that conservative messages were indeed prominent on the platform.[8, 9]

Why were critics upset to learn about Facebook's alleged editorializing? After all, it is common for websites and platforms to provide curated lists of highlighted content. One explanation is that people were concerned about whether Facebook was editing with a goal to promote a particular viewpoint without disclosing this fact,

---

[6] Mark Zuckerberg, FACEBOOK (May 12, 2016, 9:06 PM), https://www.facebook.com/zuck/posts/10102830259184701.

[7] *Trending Review Guidelines*, FACEBOOK NEWSROOM, https://fbnewsroomus.files.wordpress.com/2016/05/full-trending-review-guidelines.pdf (last visited Oct. 3, 2017).

[8] Joel Kaplan, FACEBOOK (May 14, 2016, 5:00 PM), https://www.facebook.com/joeldkaplan1/posts/1333825166634607; Brandon Silverman, *Facebook, Conservative News and How You Get to 1.5 Billion Users*, MEDIUM: THE STARTUP (May 13, 2016), https://medium.com/@brandon33175/facebook-conservative-news-and-how-you-get-to-1-5-billion-users-54a40ebbd7cd#.ibc230skr.

[9] Critique of Facebook's editorial hand is not new. In August 2014, Zeynep Tufekci argued that Facebook's News Feed was algorithmically suppressing news of the Ferguson, Missouri, protests against the police shooting of a black teenager. She wrote: "Acting through computational agency, Facebook's algorithm had 'decided' that such stories did not meet its criteria for 'relevance'—an opaque, proprietary formula that changes every week, and which can cause huge shifts in news traffic, making or breaking the success and promulgation of particular stories or even affecting whole media outlets." Zeynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 COLO. TECH. L.J. 203, 213 (2015).

a journalistic ethical debate that far predates the advent of algorithmic decisions.[10]

Of course, even without active human editorial intervention, no algorithm is fully immune from the human values of its creators. Algorithms are written by human designers, who can infuse them with their values, and are trained on human-generated data, which carry their own biases. Algorithms codify human choices about how decisions should be made. Commenting on previous controversy surrounding the "pureness" of Facebook's algorithm, scholars wrote that critics reminded them of "Captain Renault's protest as he walked into a casino in 'Casablanca': 'I'm shocked, *shocked* to find gambling in here!'"[11]

For Frank Pasquale, the episode was another example of the "black box" metaphor, depicting hidden forces embedded in algorithms to tailor and align them with corporate agendas. This Article argues that while it is true that all algorithms encapsulate human bias, the public sentiment around the "Trending Topics" story reflected a clear distinction between automated machine decision-making and more subjective social engineering. People clearly draw a distinction between opposing poles of intentional, policy-directed algorithms on the one hand and cases where bias slips into the code on the other. People want to know when they are being watched, nudged, or actively manipulated by others. There seems to be a qualitative difference between how people view automated decisions and ones that are more intentionally driven by a policy agenda.

The Facebook News Feed, which, upon its launch, was met by stiff resistance from critics and advocates, is automatically tailored for each individual based on his or her network, interests, and

---

[10] Christopher Mims, *Fears of Facebook Bias Seem to Be Overblown*, WALL STREET J. (May 16, 2016, 12:17 AM), http://www.wsj.com/articles/fears-of-facebook-bias-seem-to-be-overblown-1463371261.

[11] Jules Polonetsky & Omer Tene, *The Facebook Experiment: Gambling? In This Casino?*, RECODE (July 2, 2014, 12:55 PM), http://www.recode.net/2014/7/2/11628536/the-facebook-experiment-is-there-gambling-in-this-casino.

behavior.[12] Despite research indicating that individuals are largely averse to personalization of content and ads,[13] the News Feed has been a resounding success.[14] However, a case of undisclosed and policy-driven manipulation of the News Feed, even for a noble goal such as published academic research, has been met with consternation. People like to know if the information they are receiving has been manicured in any way.[15]

The ethics of editing algorithms mirrors the concern around native advertising, the practice of embedding paid-for ads in content such as news, product reviews, editorials, or entertainment.[16] In the native ad context, consumers need to be able to distinguish between editorial content and ads. Similarly, in algorithmically-curated environments, consumers should know when companies present them with an apparently automated but in fact edited and controlled version of reality.

While technically true, the fact emphasized by some critics that algorithms are designed artifacts, and therefore subject to human bias, may lead us to overlook an essential difference between types

---

[12] *See* Michael Arrington, *Facebook Users Revolt, Facebook Replies*, TECHCRUNCH (Sept. 6, 2006), http://techcrunch.com/2006/09/06/facebook-users-revolt-facebook-replies/ ("There has been an overwhelmingly negative public response to Facebook's launch of two new products yesterday. The products, called News Feed and Mini Feed, allow users to get a quick view of what their friends are up to . . . .").

[13] *See generally* Rena Coen, Emily Paul, Pavel Vanegas, Alethea Lange & G. S. Hans, A User-centered Perspective on Algorithmic Personalization 23–25 (May 6, 2016) (unpublished M.A. final project, University of California, Berkley), http://www.ischool.berkeley.edu/files/projects/algorithmic-personalization-coen-paul-vanegas.pdf (highlighting that respondents strongly disfavored content personalization based on race and income factors); Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley & Michael Hennessy, *Americans Reject Tailored Advertising and Three Activities That Enable It*, SSRN (Sept. 9, 2009), http://ssrn.com/abstract=1478214.

[14] James B. Stewart, *Facebook Has 50 Minutes of Your Time Each Day. It Wants More*, N.Y. TIMES (May 5, 2016) http://www.nytimes.com/2016/05/06/business/facebook-bends-the-rules-of-audience-engagement-to-its-advantage.html.

[15] *See* Turow et al., *supra* note 13.

[16] Chris Jay Hoofnagle & Eduard Meleshinsky, *Native Advertising and Endorsement: Schema, Source-based Misleadingness, and Omission of Material Facts,* TECH. SCI. (Dec. 15, 2015), http://techscience.org/a/2015121503.

of algorithms. A distinction should be drawn between policy-neutral algorithms, which provide largely unedited results that may or may not be fair or just, and edited algorithms, which are intentionally crafted to promote a policy agenda. A purist approach proves too much. It risks conflating two distinct approaches for human intervention, which are perceived differently in public opinion.

This Article addresses some of the questions raised by bias and discrimination in algorithmic decision-making. It helps separate general concerns about the digital divide and unequal access to technology from emerging dilemmas surrounding bias in algorithmic decisions. It proposes a distinction between managed, policy-directed algorithms, which companies intentionally imbue with values and norms, and policy-neutral algorithms, which, while manmade and therefore prone to bias and error, provide an unmanipulated set of results. In some cases, these results require additional adjustments after the fact to mitigate inequities or inject an editorial opinion.

This Article suggests that rather than faulting the technology of algorithms as the driver of injustices, decision makers should weigh the new technologies' risks against their formidable benefits, which include unearthing and mitigating formerly discrete and muted discrimination. Absent a comprehensive theory of discrimination, which transcends the current restrictions on bias in credit, employment, and housing norms, companies should not be expected to whitewash inequalities at the algorithmic level, lest problems in need of solutions continue to evade public scrutiny. At times, a fair policy outcome will call for editorial decisions to address inequity; at other times, it may not. Importantly, when companies do employ an editorial hand, they must be bound by requirements of transparency and accountability to avoid the specter of shadowy social engineering. The transparency called for is not one based on revealing source code, but rather public accountability about the policy-directed nature of the algorithm.

I. FAULTY ALGORITHMS; FAULTY HUMANS

Legend tells that in the late Sixteenth century, Judah Loew ben Bezalel, the rabbi of Prague, who was also known as the *Maharal*,

used clay to create a *Golem* to defend the Prague Jewish ghetto from persecution and pogroms.[17] The Golem, an animated, anthropomorphic being, was called upon in times of crisis to fight anti-Semitism, blood libel, and discrimination. Despite its strength and redeeming qualities, the Talmud considered the Golem dim-witted because, like a modern-day robot, it was literal-minded, unable to speak its mind, and lacked emotional intellect.[18] In fact, to this day, Hebrew speakers use "Golem" pejoratively to refer to a person who is daft. Opinions differ about what led to the Golem's demise. Some say it was immobilized by the Maharal himself so it would not desecrate the Sabbath; others say that it fell in love with a girl, and when rejected, became violent; yet others fear it went on a murderous rampage.[19]

Regardless of the Golem's real fate, the moral is clear: beware the human hubris and pretentiousness in trying to emulate the work of a creator. All human creation, even that of an anthropoid, must be subject to morality (and Divine law) as exemplified by the Sabbath. Failure to comply can result in disaster.

A. Faulty Algorithms

In an age of algorithmic decisions, data analytics, and artificial intelligence, life-altering decisions are increasingly handed over to manmade machines, data-crunching Golems with significant computational skills and little wit. In their book, "A Legal Theory for Autonomous Artificial Agents," Samir Chopra and Larry White write that "[a]s we increasingly interact with these artificial agents in unsupervised settings, with no human mediators, their seeming

---

[17] *See generally* MOSHE IDEL, GOLEM: JEWISH MAGICAL AND MYSTICAL TRADITIONS ON THE ARTIFICIAL ANTHROPOID (1990). In fact, there is little historical basis for attributing the making of a Golem to the Maharal of Prague. For the role of the Golem in Jewish mysticism, *see* Gershom Scholem, *The Idea of the Golem*, *in* ON THE KABBALAH AND ITS SYMBOLISM (R. Manheim trans., 1965).

[18] *See id*.

[19] *See id*.

autonomy and increasingly sophisticated functionality and behavior raises legal and philosophical questions."[20]

The big data policy debate has focused on the promises and risks of algorithmic decision-making.[21] With companies sifting through reams of data that consumers leave in their digital trails to learn new lessons and discover hidden correlations, lawyers and ethicists have argued that algorithms must be reined in.[22] They have called for transparency, equity, and fairness in automated decision-making;[23] for processes infused with values;[24] and for prevention of digital bias and discrimination.[25] The Federal Trade Commission ("FTC") asserted companies must correct for biases that could be incorporated into automated processes at both the collection and

---

[20] SAMIR CHOPRA & LARRY WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS 2 (2011).

[21] The term "big data" is typically understood to capture the steep rise in the volume of data collected and stored by business and government organizations. "The trend is driven by reduced costs of storing information and moving it around in conjunction with increased capacity to instantly analyze heaps of unstructured data using modern experimental methods, observational and longitudinal studies, and large scale simulations." Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 240 (2013); *see also* Ira S. Rubinstein, *Big Data: The End of Privacy or a New Beginning?*, 3 INT'L DATA PRIVACY L. 74 (2012); EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[22] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); *see also* Oscar H. Gandy, *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS & INFO. TECH. 29 (2008).

[23] Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 37 (2013); The Leadership Conference, *Civil Rights Principles for the Era of Big Data*, CIVILRIGHTS.ORG (2014), http://archives.civilrights.org/press/2014/civil-rights-principles-big-data.html.

[24] Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harm*, 55 B.C. L. REV. 93, 109 (2014).

[25] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U.L. REV. 1249, 1251 (2008); *see also* Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HARV. BUS. REV. (Jan. 29, 2014), https://hbr.org/2014/ 01/big-datas-dangerous-new-era-of-discrimination.

analytic stages, and "balance the predictive value of the model with fairness considerations."[26]

The press and academic literature are awash with stories and research projects demonstrating algorithmic inequities, prejudice, and bias.[27] Search engines, dating apps, e-commerce sites, and even chat bots are accused of racism, misogyny, and class discrimination. Critics require companies to cleanse algorithms, inject human discretion into decision-making, and provide due process for consumers.

B. Faulty Humans

This Article argues, however, that some of these criticisms miss the mark. On closer scrutiny, it is not clear that algorithms are to blame for many of the inequities critics identify. Moreover, re-engineering algorithms against an unsettled ethical backdrop is not necessarily the right approach to combat the moral challenges plaguing our imperfect society. In many of the most celebrated examples, the critique did not actually expose *faulty algorithms* but rather anecdotal reflections of society's deep-rooted biases and a lingering digital divide. Tweaking code and recalibrating machines may not foster fairness in the long run, but may instead sweep problems under the carpet.

To be sure, algorithms that implement discriminatory criteria are unlawful and/or unethical and must be purged. Further, algorithmic discrimination need not be direct—that is, forthrightly written into code by programmers. It could result indirectly from training algorithms on biased datasets, or using mirrors and proxies to substitute apparently benign attributes (e.g., zip code) for

---

[26] FED. TRADE COMM'N REPORT, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? (2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

[27] *See, e.g.*, Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PRO PUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Will Knight, *Biased Algorithms Are Everywhere, and No One Seems To Care*, MIT TECH. REV. (July 12, 2017), https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/.

membership in a protected class ("redlining")[28]. As Moritz Hardt explains, the "idea of 'fairness through unawareness,' . . . fails due to the existence of 'redundant encodings.' Even if a particular attribute is not present in the data, combinations of other attributes can act as a proxy."[29] Algorithmic parameters are never neutral. They are always imbued with values. As Anna Lauren Hoffman reminds us, "ontologies are not born out of nothing. Instead, they emerge from (and are shaped by!) the active, open-ended, and everyday practices of the world they purport to describe."[30]

But in cases where legal requirements have been met, modifying algorithms to correct for disparities may cleanse the public space, and create a manicured environment, without jettisoning underlying societal biases. Worse, such under-the-hood tampering could simply replace the prejudices and biases of a divided polity with those of Silicon Valley engineers and entrepreneurs, whom themselves have been accused of perpetuating a white male-centric environment.[31]

If laws and norms fail to adequately cure social inequities, policymakers should work to change those laws and norms, rather than burying them under a gloss of binary code.[32] When companies

---

[28] Jessica Silver-Greenberg, *New York Accuses Evans Bank of Redlining*, N.Y. TIMES (Sept. 2, 2014, 12:01 AM), https://dealbook.nytimes.com/2014/09/02/new-york-set-to-accuse-evans-bank-of-redlining/.

[29] Moritz Hardt, *Equality of Opportunity in Machine Learning*, GOOGLE RESEARCH BLOG (Oct. 07, 2016), https://research.googleblog.com/2016/10/equality-of-opportunity-in-machine.html.

[30] Anna Lauren Hoffmann, *Science Will Not Save Us: Medicine, Research Ethics, and My Transgender Body*, AUTOSTRADDLE (July 16, 2014, 9:00 AM), http://www.autostraddle.com/science-will-not-save-us-medicine-research-ethics-and-my-transgender-body-240296/.

[31] *See, e.g.*, Dominic Rushe, *Twitter's Diversity Report: White, Male and Just Like the Rest of Silicon Valley*, THE GUARDIAN (July 25, 2014, 11:16 AM), http://www.theguardian.com/technology/2014/jul/25/twitter-diversity-white-men-facebook-silicon-valley; David Streitfeld, *Ellen Pao Loses Silicon Valley Bias Case Against Kleiner Perkins*, N.Y. TIMES (March 27, 2015), http://www.nytimes.com/2015/03/28/technology/ellen-pao-kleiner-perkins-case-decision.html.

[32] *Cf.* Julie Brill, Commissioner, Fed. Trade Comm'n, Keynote Address Before Coalition for Networked Information Fall 2015 Membership Meeting: Transparency, Trust, and Consumer Protection in a Complex World (Dec. 15, 2015),

decide to tame the Golem by censoring offensive content and prejudiced results, thereby substituting their values and ideals for those of a broader user base, they must provide transparency into their decisions and actions. Without transparency, critics could accuse companies of digital manipulation, narrowcasting, or even swaying elections.[33]

C. Is It the Algorithm?

When addressing accusations of algorithmic bias, organizations should first assess whether the purported bias is endogenous to the algorithm or overlaid by the actions of human actors. In many cases, criticisms of algorithmic decisions in fact reflect broader concerns about a digital divide or even a general condemnation of an unequal society.[34] For example, as shown below, concerns over bias on leading sharing economy services such as Airbnb and Uber have less to do with these services' algorithms and more with how a biased customer base makes use of them.[35] Attention should therefore focus on new situations where algorithms are the primary driving force. In these cases, algorithms are used to make decisions about credit or job applications, an advertisement served, or social connection made. Where algorithms are used to churn through piles of data to make life-altering decisions, give recommendations, deliver content, or provide authoritative responses, taming the Golem becomes salient.

D. Policy-neutral vs. Policy-directed Algorithms

Where potential bias is endogenous to the algorithmic process, it is useful to distinguish between two categories of algorithms. *The*

---

https://www.ftc.gov/system/files/documents/public_statements/895843/151216cnikeynote.pdf (discussing algorithmic accountability for discrete discrimination and concluding, "[u]ltimately, I believe we need legislation to address many of these issues.").

[33] Jonathan Zittrain, *Engineering an Election: Digital Gerrymandering Poses a Threat to Democracy*, 127 HARV. L. REV. F. 335 (2014).

[34] *Cf.* COUNCIL OF ECON. ADVISORS, MAPPING THE DIGITAL DIVIDE (July 2015), https://www.whitehouse.gov/sites/default/files/wh_digital_divide_issue_brief.pdf.

[35] *See infra* Section II.B.2.

first category, which we call "*policy-neutral algorithms*," comprises algorithmic processes that are largely expected to provide a neutral, objective, mathematical result.[36] What is the most profitable location for a new business? Which result do users click on when they search for the word "Jew"? Here, users would be surprised to discover they are being presented a manicured, edited vision of the world.

Raw mathematical calculations should not be tailored to fit what a designer views as just, fair, or politically correct. To be sure, even policy-neutral algorithms are manmade, written by engineers and trained on selected datasets that could reflect inequity or bias. Some degree of editing is required to make the results of any algorithm clear, readable, and coherent. Yet a bright line crosses between such instances of micro-level engineering of source code and cases where algorithms are intentionally imbued with a designer's values and norms.

Requiring untainted calculations does not prevent remedial action or social engineering based on the lessons learned from policy-neutral algorithms. A retailer may decide to open a store in a poor neighborhood to include the local population despite an algorithmic process that predicts low-profit margins. A search engine may add a notice to explain why search results for the word "Jew" are filled with hate speech. But the output of the algorithm itself remains unaltered.

The second category, which we call "*policy-directed algorithms*," comprises algorithms used as input toward intentional top-down editorial or policy directed choices. For example, a matchmaking app directs a user to experiment with dating users from different faiths not because it predicts the highest likelihood of

---

[36] Gilad Lotan suggests "a distinction between 'supervised' and 'unsupervised'" algorithms. The latter involves "letting the pattern speak for itself," while "the former requires taking specific data and drawing from it to achieve a specific objective." DATA & SOC'Y RESEARCH INST., WHO CONTROLS THE PUBLIC SPHERE IN AN ERA OF ALGORITHMS 1, 5 (2016), http://www.datasociety.net/pubs/ap/WorkshopNotes_PublicSphere_2016.pdf. We use "policy-neutral" versus "policy-directed" here, recognizing that even unsupervised algorithms are supervised and conveying the deeper involvement of human editorializing in the policy directed form.

success, but rather because the designer is promoting a policy of equity and fairness.

In these cases, transparency is key. Not transparency around the source code or inner workings of the algorithm, which are clearly indecipherable for the general public, but rather about the fact that the process is actively managed and edited, and in certain cases, about the principles and policies advanced by the curator-editor.[37] Just as Facebook recently concluded before disclosing its editors' guide, transparency is the best remedy for allegations of political tinkering. Transparency means algorithmic editing must be visible, reviewable, and expressly declared. Individuals should know whether the search results and content they see are policy-neutral, or, conversely, reflective of a curator's mission. Even when intentions are noble, curators could get in trouble for discreetly advancing an agenda without adequate transparency and internal oversight. The stakes are even higher in nondemocratic societies, where governments can maneuver public perceptions and behavior via surreptitious editing of algorithmic code.

Individuals have a right to know whether a digital product or service they are offered operates as, to use Jack Balkin's term, an "information fiduciary," or as a tool to advance a corporate or larger societal agenda.[38] "Is my wearable device serving me or secretly enrolling me in a social experiment?": this is a question any user would want answered. In a similar vein, the FTC's policy statement on native advertising explains that:

> an ad's format is deceptive if it materially misleads consumers about the ad's commercial nature, including through any implied or express representation that it comes from a party other than the sponsoring advertiser. If the source of advertising content is clear, consumers can make informed decisions about whether to interact with the advertising and the weight to give the information conveyed in the ad.[39]

---

[37] Brill, *supra* note 32.

[38] *See* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183, 1186 (2016).

[39] Press Release, Fed. Trade Comm'n, FTC Issues Enforcement Policy Statement Addressing "Native" Advertising and Deceptively Formatted Advertisements (Dec. 22, 2015), https://www.ftc.gov/news-events/press-releases/2015/12/ftc-issues-

In Europe, the new European General Data Protection Regulation ("GDPR") provides that "[e]very data subject should . . . have the right to know and obtain communication in particular with regard to . . . the logic involved in any automatic personal data processing and, at least when based on profiling, the consequences of such processing."[40]

Imagine, for example, a navigation app that diverts a user to a longer route home in order to minimize traffic congestion in the city instead of directing the driver to take the quickest ride. Such a service may advance important societal goals such as minimizing congestion and reducing emissions, but, unless it clearly disclosed its motivations, it would risk alienating users and losing their trust. Or, consider an urban app that purports to help tourists find the safest streets and neighborhoods in a city, but it instead sends them to crime-ridden areas in order to correct for social inequities. Here too, the general public policy agenda may be at odds with the interest of a specific user, who should at least be aware he or she is nudged by a paternalistic force.

In practice, given that policy-neutral algorithms are edited—indeed written—by humans, the distinction between such algorithms and policy-directed algorithms is not always crisp. For example, a content provider may edit a policy-neutral list of algorithmically selected news items to prevent repetition, or an editor may add short written summaries to each item, thereby introducing policy bias. Rather than being a dichotomy, algorithms lie on a spectrum with policy-neutral and policy-directed algorithms marking the two poles.

The distinction between consumers' perception of information fiduciaries and companies pursuing their own agenda can help

---

enforcement-policy-statement-addressing-native; *see also* FED. TRADE COMM'N, COMMISSION ENFORCEMENT POLICY STATEMENT ON DECEPTIVELY FORMATTED ADVERTISEMENTS (2015), https://www.ftc.gov/system/files/documents/public_statements/896923/151222deceptiveenforcement.pdf.

[40] Regulation (EU) 2016/679, of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1, 12.

explain the trust gap between explicit personalization services of companies such as Amazon ("If you liked this you may also like that") and more opaque personalization by ad serving companies. When polled, many Americans consistently express negative sentiment toward the model of online personalization despite its clear benefits in terms of relevant content and ads, not to mention the support of an economic model of "free."[41] Conversely, for companies viewed as information fiduciaries, customization is embraced. Few if any consumers are upset by Amazon's personalized offerings given the company's clear messaging, user-friendly interface, and general brand recognition for targeted marketing. At the same time, ad tech companies are perceived as pursuing their own agenda, setting algorithms to maximize ad inventory value, and monetizing with little apparent value for consumers, and are therefore met with suspicion and, increasingly, ad-blocking tools.[42]

When implementing policy-directed algorithms, designers must follow processes for setting forth, reviewing, and auditing ethical standards and rules. For many years, traditional media organizations such as the *New York Times* have operated under assumptions that editors had a responsibility to use honest judgment to determine which stories to cover and how to highlight them. Television networks implemented Broadcast Standards and Practices to account for moral, ethical, and legal implications of programs and advertisements they aired.[43] Although the click-driven priorities of

---

[41] TUROW ET AL., *supra* note 13, at 24; *see also* Chris Jay Hoofnagle & Jennifer M. Urban, *Alan Westin's Privacy Homo Economicus*, 49 WAKE FOREST L. REV. 261 (2014); Chris Jay Hoofnagle et al., *Behavioral Advertising: The Offer You Cannot Refuse*, 6 HARV. L. & POL'Y REV. 273 (2012); JOSEPH TUROW ET AL., AMERICANS, MARKETERS, AND THE INTERNET: 1999–2012 (2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2423753. *See generally* Chris Jay Hoofnagle & Jan Whittington, *Free: Accounting for the Costs of the Internet's Most Popular Price*, 61 UCLA L. REV. 606 (2014).

[42] *See generally* JOSEPH TUROW, THE DAILY YOU: HOW THE NEW ADVERTISING INDUSTRY IS DEFINING YOUR IDENTITY AND YOUR WORTH (2011) (describing the emergence of the retail market as a heavily monitored surveillance environment driven by targeted advertising).

[43] *See, e.g.*, NBC UNIVERSAL, ADVERTISING GUIDELINES 14 (2017), https://www.nbcuadstandards.com/files/NBC_Network_Advertising_Guidelines.pdf.

many new media outlets seem to overwhelm any sense of curation, publishers that are perceived as editorial in nature are expected to be accountable for the policy decisions behind the content and types of ads they display.

E. Editorial Dilemmas

Implementing editorial discretion and social engineering raises a broad array of ethical dilemmas. Are all companies well placed to identify societal consensus and norms, or should they be trailblazers adopting progressive agendas? With the concern for the pervasiveness of "fake news" on social media before the elections, would society be better off with companies such as Facebook or Google actively policing the Internet for untruthful content? Even setting aside the daunting operational challenge of sifting through endless amounts of information in real time, are companies well placed to act as arbiters of the truth? If so, under which cultural, normative, and ideological standard do we expect them to act?

An approach requiring companies to "tame the Golem," bringing algorithmic decisions to heel by instilling them with liberal values, is based on fragile grounds. First, such an approach could incentivize companies to sweep socially fraught issues under the carpet, sanitizing decisions to present users with a *Shallow Hal* view of the world.[44] If Google suppressed hateful search results for the word "Jew," the underlying social problems would not be solved, but rather concealed from public view.

Second, it places business entities, which are undemocratic bureaucracies with little transparency, due process, or accountability, in the unenviable position of being the final arbiters of ethical dilemmas and social norms.[45] Corporations are legal

[44] In the movie *Shallow Hal*, Shallow Hal is hypnotized into only seeing a person's inner beauty. When the spell breaks, he is forced to confront a more flawed version of the world. *See* SHALLOW HAL (Twentieth Century Fox 2001).

[45] *See generally* Jeffrey Rosen, *The Deciders: The Future of Privacy and Free Speech in the Age of Facebook and Google*, 80 FORDHAM L. REV. 1525 (2012) (discussing the weighty policy decisions that befall officers at online platforms as they navigate foreign laws, regulations, and cultural sensitivities while at the same time trying to satisfy American values such as freedom of speech).

constructs intended to maximize profit and shareholder value. Many do not have ethics review processes, chief privacy officers, or other mechanisms for arbitrating social values and norms. A prime example is a case from the European Court of Justice establishing a "right to be forgotten,"[46] which effectively charged Google with balancing delicate values, norms, and fundamental rights including freedom of speech, freedom of information, and the right to privacy, and doing so in a variety of cultural and legal environments all over the world.[47] Although privacy advocates claimed victory, critics argued that at the end of the day the decision endowed the company with tremendous discretion with little legal guidance.[48]

Third, when viewed by users from other countries and cultures in Asia, Africa, and beyond, resolution of these issues under Silicon Valley ethics could be considered American-centric, socio-technological colonialism, thus imposing Western liberal values on societies that have broadly divergent views about gender, family, religion, and politics.[49] Even within the U.S., as the election results clearly demonstrated, tech leaders may be out of sync with popular values. Advocates of proactive corporate editorializing should bear in mind that this approach could cut both ways. For example, in the recent debate over North Carolina's legislation proscribing transgender individuals' access to public bathrooms,[50] companies such as Target and Bank of America led the charge for more liberal

---

[46] Case C-131/12, Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costejo Gonzalez, 2014 E.C.R. 317, at ¶¶ 89–99 (May 13, 2014), http://curia.europa.eu/juris/document/document_print.jsf?doclang=EN&docid=152065.

[47] *See* Mark Scott, *Europe Tried to Rein in Google. It Backfired.*, N.Y. TIMES (Apr. 18, 2016), http://www.nytimes.com/2016/04/19/technology/google-europe-privacy-watchdog.html?ref=technology.

[48] *See id*.

[49] *See* Jeffrey Rosen, *Google's Gatekeepers*, N.Y. TIMES MAG. (Nov. 28, 2008), http://www.nytimes.com/2008/11/30/magazine/30google-t.html (describing the challenges of Google officers as they make policy decisions about content displayed in countries like China and Turkey).

[50] H.B. 2, 2016 Gen. Assemb., 2d Extra Sess. (N.C. 2016).

laws.[51] But in other cases, companies have promoted a conservative agenda out of sync with civil rights activists who argue for additional involvement.[52]

Cleansing algorithms of undesirable values is equivalent to policies nudging individuals toward better outcomes for themselves (e.g., a wearable device reminding a user to exercise) or society at large (e.g., a social network nudging users to increase voter turnout).[53] Critics have long argued that nudging, also known as Libertarian Paternalism, is merely paternalism in disguise, encouraging abuse of power by technocrats and impairing individuals' autonomy to make moral choices.[54] The outrage unleashed by New York City's attempt to nudge consumers to reduce intake of sugary soda by banning its sale in large cups is a case in point, forewarning companies against surreptitious meddling with policy neutral algorithms.[55] As evident in the public storm around the Facebook emotional contagion study, digital manipulation, even if undertaken for a noble cause, such as research

---

[51] *See, e.g.*, Barb Darrow, *Bank of America Joins Fight Against North Carolina Bathroom Law*, FORTUNE (Mar. 30, 2016, 8:33 AM), http://fortune.com/2016/03/30/boa-pushes-slams-north-carolina-bathroom-law; Robert Mclean, *Target Takes Stand on Transgender Bathroom Controversy*, CNN (April 20, 2016, 2:26 AM), http://money.cnn.com/2016/04/20/news/companies/target-transgender-bathroom-lgbt/index.html.

[52] *See, e.g.*, Burwell v. Hobby Lobby, 134 S. Ct. 2751 (2014) (upholding a company's owners right to run the business according to their religious faith, including the belief that the use of contraception is immoral). This demonstrates that deferring to corporate driven normative policies will not necessarily result in a liberal-leaning agenda.

[53] *See* Richard H. Thaler & Cass R. Sunstein, *Libertarian Paternalism*, 93 AMERICAN ECON. REV. 175 (2003). *See generally* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (Revised and Expanded ed., Penguin Books 2009) (2008).

[54] *See* Gregory Mitchell, *Libertarian Paternalism Is an Oxymoron*, 99 NW. U. L. REV. 1245 (2005); *cf.* Richard H. Thaler & Cass R. Sunstein, *Libertarian Paternalism Is Not an Oxymoron*, 70 U. CHI. L. REV. 1159 (2003).

[55] *See* Steven J. Gonzalez, *Assisting Personal Responsibility: Using Nudges to Reduce Sugar Consumption*, HARV. L. & POL'Y REV.: ONLINE PIECES (Mar. 17, 2017), http://harvardlpr.com/2017/03/17/assisting-personal-responsibility-using-nudges-to-reduce-sugar-consumption.

to determine whether a claimed product defect actually existed and to advance generalizable knowledge, can marshal a forceful consumer backlash.[56]

Transparency around policy-directed algorithms is key not only in the field of online content and social networks but also in the burgeoning Internet of Things. Here, engineers face an immediate necessity to—one way or another—translate ethics into code. For example, the trolley problem, a thought experiment often discussed in introductory philosophy courses, has become an immediate operational dilemma for designers of autonomous vehicles.[57] How should a vehicle act when a crash is inevitable and lives are at stake? Should it account for factors such as life expectancy, earning potential, genetic disposition to disease and insurance coverage of the drivers, passengers and pedestrians involved?

With cyber-physical system design, the stakes are often higher than with content platforms. It is one thing to be nudged into reading a news report or seeing an ad; it is quite another thing to be surreptitiously manipulated in ones' offline daily activities. A user has a right to know if their smart car sets them on a course they did not expect in order to fulfill a social agenda, or if their smart thermostat lowers the temperature in their house to conserve public resources. Social values in design may well reflect important policy goals, but regardless of the merits of nudging by government and businesses, individuals have a right to know if the reality they experience is contrived.

Legal requirements or social justice may call for companies to make essential policy decisions that defy the automated choices of

---

[56] *See* Adam D. I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental Evidence of Massive-scale Emotional Contagion Through Social Networks*, 111 PROC. NAT'L ACAD. SCI. 8788 (2014), http://www.pnas.org/content/111/24/8788.full.pdf; *see also* Vindu Goel, *Facebook Tinkers with Users' Emotions in News Feed Experiment, Stirring Outcry*, N.Y. TIMES (June 29, 2014), http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html.

[57] *See* Joel Achenbach, *Driverless Cars Are Colliding with the Creepy Trolley Problem*, WASH. POST (Dec. 29, 2015), https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/?utm_term=.ac26446f5e36.

algorithmic analysis. However, as discussed below, these changes should either be cast as policy decisions made despite the output of algorithms or transparently disclosed as editorial decisions so that users understand the nature of the services offered.

II. CASE STUDIES

This part analyzes a series of case studies reported in the press or academic literature as examples of algorithmic discrimination and bias. It highlights the distinctions set forth above between cases that genuinely feature algorithmic decision-making and others that are better characterized as technologically mediated human decisions. With respect to algorithmic decision-making cases, it distinguishes between policy-neutral algorithms and policy-directed algorithms, suggesting a need for full disclosure of policies and practices with respect to editorial choices.

A. Algorithmic Decision-making

This section provides examples of cases characterized by algorithmic decision-making. It demonstrates that in such cases, a distinction should be drawn between policy-neutral algorithms, which deliver unedited results, and policy-directed algorithms, which serve the designer's policy agenda.

1. Search and Ads

In 2013, Harvard computer science professor Latanya Sweeney conducted a series of research experiments on the contextual ads placed next to various searches of individuals' names.[58] Sweeney discovered that when searching Google for a "black-sounding" name, such as DeShawn, Darnell, or Jermaine, ad results included references to arrest or criminal records at a significantly higher rate than when searching for traditionally "white-sounding" names, such as Geoffrey, Jill, or Emma.[59]

---

[58] *See* Latanya Sweeney, *Discrimination in Online Ad Delivery*, COMM. ACM, May 2013, at 44, https://cacm.acm.org/magazines/2013/5/163753-discrimination-in-online-ad-delivery/fulltext.

[59] *Id*. at 46–47.

Reporting these findings, the *MIT Technology Review* accused, "Racism is Poisoning Online Ad Delivery,"[60] while the *Huffington Post* similarly wrote "Google's Online Ad Results Guilty of Racial Profiling."[61] Sweeney conjectured that the fact that black-sounding names were more likely to yield such advertisements resulted from the algorithmic process that Google employs to determine which advertisements to place.[62] While proprietary, the algorithms known to rely on the tendency of users to actually click on an ad when it makes placement decisions. Over time, as people click one version of an ad more often than others, the weights assigned by the algorithm change, and the ad text getting the most clicks eventually displays more frequently. The differential delivery of ads therefore reflected the prejudice already held by the users who were exposed to the ads. In other words, Google's ad delivery service was a mirror placed in front of a biased society.[63]

Sweeney's research revealed that one party in particular, a data broker named "Instant Checkmate," gamed the system by bidding highest for the term "arrest record" and attaching it to any name in its directory.[64] The company was thus able to win Google ad auctions for any search related to arrest records of specific individuals.[65] Apparently, the results Sweeney reported were brought on by the higher incidence of users clicking on "arrest

---

[60] *Racism Is Poisoning Online Ad Delivery, Says Harvard Professor*, MIT TECH. REV. (Feb. 4, 2013), https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/.

[61] Bianca Bosker, *Google's Online Ad Results Guilty of Racial Profiling, According to New Study*, HUFFINGTON POST (Feb. 13, 2013, 11:14 AM), http://www.huffingtonpost.com/2013/02/05/online-racial-profiling_n_2622556.html.

[62] *Racism Is Poisoning Online Ad Delivery, Says Harvard Professor*, *supra* note 60.

[63] *See generally* LAWRENCE PAGE, SERGEY BRIN, RAJEEV MOTWANI & TERRY WINOGRAD, THE PAGERANK CITATION RANKING: BRINGING ORDER TO THE WEB, STANFORD INFOLAB (1998), http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

[64] *Racism Is Poisoning Online Ad Delivery, Says Harvard Professor*, *supra* note 60.

[65] For an explanation of Google ad auctions, see GOOGLE ADWORDS, https://adwords.google.com.

record" ads in conjunction with black sounding names. "Instant Checkmate" was later the target of an FTC enforcement action in connection with its business practices.[66]

Additional work demonstrated apparent bias in the placement of Google ads. A group of Carnegie Mellon researchers recently demonstrated that after searching Google for available job openings, male job seekers were much more likely than equivalent female job seekers to be shown ads for high-paying executive jobs.[67] In contrast, when researchers set the gender of user agents to female, fewer instances of ads related to high-paying jobs were shown.[68] For example, one experiment demonstrated that Google displayed ads for a career coaching service for "$200k+" executive jobs 1,852 times to a male user agent group, compared to just 318 times to an equivalent group of women.[69]

Sweeney's research is one of various reported cases where Google's search algorithm delivered results that reflected the bigotry and prejudice of a divided public. Unlike organic search, Google's ad environment is an example of a *policy-directed algorithm*. Users do not expect Google's ad delivery platform to serve unaltered objective truths. Accordingly, Google and other advertising platforms establish ethical standards and practices to promote a more ethical environment.[70] In our view, advertisers should not be allowed to commission or display hateful ads or bid on search terms that are manifestly racist or criminal. Nor should

---

[66] *See* Press Release, Fed. Trade Comm'n, Two Data Brokers Settle FTC Charges That They Sold Consumer Data Without Complying with Protections Required Under the Fair Credit Reporting Act (Apr. 9, 2014), https://www.ftc.gov/news-events/press-releases/2014/04/two-data-brokers-settle-ftc-charges-they-sold-consumer-data.

[67] *See* Amit Datta, Michael Carl Tschantz & Anupam Datta, *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*, ARXIV.ORG, (Mar. 18, 2015), https://arxiv.org/pdf/1408.6491.pdf.

[68] *Id.*

[69] *Id.* at 13.

[70] *See, e.g.*, GOOGLE ADWORDS POLICIES, https://support.google.com/adwordspolicy/answer/6008942?hl=en#res (last visited Sept. 21, 2017); YAHOO ADVERTISING POLICIES, https://adspecs.yahoo.com/pages/yahooadpolicies/?.tsrc=rtlde/%3fformat=rss (last visited Sept. 21, 2017).

advertisers abuse the system by pursuing and rewarding foul conjecture.

For example, Google has recently announced it would ban all advertisements for payday loans, defined by the company as "ads for loans where repayment is due within 60 days of the date of issue," as well as "ads for loans with an APR of 36% or higher" because such practices often lead to unaffordable repayment terms and financial harm to borrowers.[71] In doing so, it made an editorial decision to alter the inputs to its algorithm to advance a policy goal.[72]

Organic search results are different. Here, Google uses a *policy-neutral algorithm*, in accordance with users' expectations to obtain unedited search results. Accordingly, Google has typically chosen not to intervene to sugarcoat reality by whitewashing organic search results. For example, when users searched for the word "Jew" and obtained results linking to hate groups, presumably reflecting what users who searched for that term tended to click on, Google did not alter the search algorithm.[73] Different results would come up in searches for terms like "Jewish" or "Judaism," reflecting the

---

[71] David Gradd, *An Update to Our AdWords Policy on Lending Products*, GOOGLE (May 11, 2016), http://googlepublicpolicy.blogspot.com/2016/05/an-update-to-our-adwords-policy-on.html. This followed Facebook's longstanding policy. *See What Are Facebook's Advertising Policies Around Financial Loan Companies?*, https://www.facebook.com/policies/ads (last visited Sept. 21, 2017); Francine McKenna, *Google Follows Facebook in Banning Payday Loan Ads*, MARKETWATCH (May 13, 2016, 5:08 PM), http://www.marketwatch.com/story/google-follows-facebook-in-banning-payday-loan-ads-2016-05-11.

[72] Critics argued that even the decision to exclude ads for payday loans, apparently benefitting from broad consensus given the risk and hardship such practices impose on borrowers, are not above the fray. A representative of the Online Lenders Alliance, an association representing online financial services companies and their customers, warned that "[t]he Federal Reserve Board noted last year that 47 percent of Americans are not prepared to handle a $400 unexpected expense[.] . . . This is yet another tactic that further limits the ability of families to have access to credit to fulfill their financial obligations." Christine Hauser, *Google To Ban All Payday Loan Ads*, N.Y. TIMES (May 11, 2016), http://www.nytimes.com/2016/05/12/business/google-to-ban-all-payday-loan-ads.html. Other search engines continue to accept such loans.

[73] David Becker, *Google Caught in Anti-Semitism Flap*, CNET (Apr. 9, 2004 12:23 PM), http://news.cnet.com/2100-1038_3-5186012.html.

divergent nomenclature of different user groups. Google left the offensive search results for "Jew" intact, providing a disclosure at the top of the page explaining the reason for the offensive (but accurate) results.

The *Guardian* recently reported apparent prejudice reflected in Google's organic search results. According to the report, searching Google for the phrase "unprofessional hairstyles for work" yielded image results of mainly black women with natural hair, while searching for "professional hairstyles" offered pictures of coiffed, white women.[74] Clearly, Google searches mirror users' and publishers' deep-seated societal biases; as such, they could have an amplifying effect. But should the company bear responsibility to conceal these responses or should it reveal the bias it finds on the web?

When a user queries Google for, say "three black guys," what is he or she asking? What is reflected in content on the Internet? What is authoritative? What is right? What does Google think is (or should be) right? Should Google do anything to eliminate prejudices and disparities that are reflected back by a policy-neutral algorithm? Should it put a thumb on the scale to provide more just results?

Of course, it is morally reprehensible that anyone would think that an Afro is less "professional" than a ponytail. But unilaterally adjusting for such bias—and endless other potential biases—could easily be considered surreptitious manipulation of social values by a Silicon Valley firm, whose principles and morals do not necessarily reflect those of a broader polity. Most Google users likely expect the search engine to deliver to them an accurate representation of the information that is available online, as unpleasant as it may sometimes be. Consider, again, the public storm around the reports that Facebook doctored the "Trending Topics" side bar.[75] The prospect of surreptitious editing hits a raw

---

[74] Leigh Alexander, *Do Google's 'Unprofessional Hair' Results Show It Is Racist?*, THE GUARDIAN (Apr. 8, 2016, 3:50 AM), https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-.

[75] *See* Nunez, *supra* note 4.

nerve for users who had expected results to be assembled by an automated policy-neutral algorithm.

2. Dating Apps

A *BuzzFeed* story recently demonstrated that users of a dating app, *Coffee Meets Bagel*, who checked "no preference" for desired ethnicity of prospective dates were nevertheless automatically matched with people of their own race.[76] The algorithm, it seems, was more race-minded than its users. Explaining the biased result, representatives for the app offered, "we do so because our data shows even though users may say they have no preference, they still (subconsciously or otherwise) prefer folks who match their own ethnicity."[77] The developers argued that rather than imposing a racist—or for that matter, any other—agenda, the algorithm simply optimized for the success of the dates it set up, training on a rich dataset that apparently proved that same-race dates had a higher probability of success.[78]

Should a dating app continue using a policy-neutral algorithm or create a colorblind, policy-directed algorithm? Critics may argue that the algorithm should at least be curated to avoid favoring same-race dates more so than the app's users. Yet, if an algorithm has evidence about users' real preferences based on a large pool of results, should it be programmed to ignore those facts to act in a way that is race blind and thus more socially responsible? And if so, should it make matches that intentionally counteract existing bias, a

---

[76] Katie Notopoulos, *The Dating App That Knows You Secretly Aren't into Guys from Other Races*, BUZZFEED (Jan. 14, 2016, 10:44 AM), http://www.buzzfeed.com/katienotopoulos/coffee-meets-bagel-racial-preferences#.fh7l5zKPb2.

[77] *Id.*

[78] *See id.* Previous research by dating sites like *Ok Cupid* found broad disparities in the appeal of users from different racial backgrounds. Kat Chow & Elise Hu, *Odds Favor White Men, Asian Women on Dating App*, NPR: CODESWITCH (Nov. 30, 2013, 8:00 AM) http://www.npr.org/sections/codeswitch/2013/11/30/247530095/are-you-interested-dating-odds-favor-white-men-asian-women; Christian Rudder, *Race and Attraction, 2009–2014*, OKCUPID BLOG (Sept. 10, 2014), http://blog.okcupid.com/index.php/race-attraction-2009-2014/.

form of dating affirmative action, or perhaps just ones that are "colorblind" but may end up accentuating lingering disparities? In these cases, the law provides no guidance; dating sites are not regulated by equal opportunity legislation. And deciding which course of action is desirable and ethical can pit critics' lofty ideals against the harsh reality of the markets, for apps and dates.

3. AI, Bots, and Digital Assistants

In many parts of the market, social norms trail the rapid evolution of new technologies.[79] One such area is the fast-burgeoning field of bots, chat bots, and digital assistants.[80] Where artificial intelligence ("AI") is concerned, machines no longer only reflect the bias and prejudice of large pools of users, but rather they are required to manage human-like interactions—and make ethical decisions—on their own. Mishandling these interactions can result in embarrassing gaffes; though getting them right risks the creepy feeling associated with the uncanny valley.[81] It may be easy enough to program a robot to pace on a sidewalk without crashing into windows or bumping into people. However, engineers will inevitably have to infuse robots with human norms of behavior and etiquette—to turn their face to the door when standing in an elevator or disable their camera when stepping into a dressing room.

Recent reports have shown that bots and digital assistants, such as Apple's Siri and Microsoft's Cortana, repeatedly encountered socially inappropriate questions and comments. CNN reports, "[a]

---

[79] Omer Tene & Jules Polonetsky, *A Theory of Creepy: Technology, Privacy and Shifting Social Norms*, 16 YALE J.L. & TECH. 59, 61 (2013).

[80] *Bots, the Next Frontier*, ECONOMIST (Apr. 9, 2016), http://www.economist.com/news/business-and-finance/21696477-market-apps-maturing-now-one-text-based-services-or-chatbots-looks-poised/.

[81] In his classic 1970 article, *The Uncanny Valley*, Tokyo Institute of Technology robotics professor Masahiro Mori envisioned that people's reactions to robots that looked and acted almost human would shift from empathy to revulsion as they approached, but failed to attain, a lifelike human appearance. This descent into eeriness is known as the uncanny valley. Masahiro Mori, *The Uncanny Valley*, 7 ENERGY 33 (1970) (Japan), *translated in* IEEE ROBOTICS & AUTOMATION, June 2012, at 98, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6213238.

side effect of creating friendly female personalities is that people also want to talk dirty, confess their love, role play, or bombard them with insults."[82] How should bot designers deal with "digital sexual harassment"?

Due to the human-like aspects of chat bots, users expect them to have a point of view that is editorial in nature. Consequently, companies that design policy-directed algorithms to guide the behavior of chat bots are already facing novel ethical dilemmas. Microsoft, for example, decided that when facing digital sexual harassment, its digital assistant, Cortana, who speaks with a female voice, should stand up for its rights.[83] But other designers noted, "there is a high demand for an assistant personality that's 'more intimate-slash-submissive with sexual undertones.'"[84] How should digital assistants react to centuries of male bias and sexism? Should companies pair-up male users with male bots even against their stated preferences? Should they enforce a 50/50 split between female and male bots?

As long as machines churn through policy-neutral algorithms, simply reflecting back public sentiments, the responsibility of the designers and engineers is limited. Yet once a machine is expected to express *its own opinion* and make conscientious choices, ethical mores become salient. If asked to provide advice on available options for abortion, should a bot heed the law of the state it responds in? Should it do so even if that law prohibits abortion in cases of rape or incest? What are the correct political opinions for a bot? The *Washington Post* reports,

> [t]o field increasingly common questions about whether Cortana is a fan of Hillary Clinton's, for instance, or Donald Trump's, the team dug into the backstory to find an answer that felt 'authentic.' . . . So Cortana says

[82] Heather Kelly, *Even Virtual Assistants Are Sexually Harassed*, CNN (Feb. 5, 2016, 10:41 AM), http://money.cnn.com/2016/02/05/technology/virtual-assistants-sexual-harassment/index.html.

[83] *See id.*

[84] *Id.*

> that all politicians are heroes and villains. She declines to say she favors a specific candidate.[85]

With companies designing bots to emulate humanity, the dilemmas bots face will inevitably become more complex. If a user expresses suicidal thoughts, should a bot blow the whistle and call the police? Their spouse? Their ex-spouse? Should a bot alert social services if it detects hostility in the tone that a parent communicates with a child?

Microsoft has recently decommissioned its bot, Tay, after it was taught by users to spew racist slurs, declaring on Twitter, "Hitler was right, I hate the Jews."[86] Nazi opinions aside, which part of society should humanized bots be programmed to reflect? Trump supporters, Silicon Valley entrepreneurs, or Boston Brahmins? Indeed, should bots reflect American values at all, as opposed to Japanese, Chinese, Saudi Arabian, or Indonesian?

There are likely to be algorithmic decisions that will be hard to pigeonhole as policy neutral or editorial, for example the safety decisions that an autonomous vehicle will have to make in navigating the proverbial trolley problem and other ethically complex decisions.

4. Object Recognition

An additional challenge in taming the Golem requires teaching it to develop the emotional intelligence and moral aptitude to know not only which decisions to make, but also which mistakes to avoid. Last year, Google had to publicly apologize after its object recognition algorithm tagged two black users of Google Photo as "gorillas."[87] In this case, Google's object recognition algorithm did

---

[85] Elizabeth Dwoskin, *The Next Hot Job in Silicon Valley Is for Poets*, WASH. POST: SWITCH (Apr. 7, 2016), https://www.washingtonpost.com/news/the-switch/wp/2016/04/07/why-poets-are-flocking-to-silicon-valley.

[86] Peter Bright, *Microsoft Terminates Its Tay AI Chatbot After She Turns into a Nazi*, ARS TECHNICA (Mar 24, 2016, 10:28 AM), http://arstechnica.com/information-technology/2016/03/microsoft-terminates-its-tay-ai-chatbot-after-she-turns-into-a-nazi/.

[87] Alistair Barr, *Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms*, WALL ST. J.: DIGITS (July 1, 2015, 3:40 PM), http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/.

not mis-tag the users because it was racist; rather, it did so because it *was not* racist, and lacked the emotional wherewithal to understand that such a decision is socially charged. The bot's mistake is imbued with racist overtones only when humans read their own prejudice into the situation. Tagging a black person as a gorilla has emotional baggage for humans, not for bots. In his piece, *Why Robots Can't Become Racist, and Why Humans Can*, Matthew Nowachek explains,

> robots cannot become racist insofar as their ontology does not allow for an adequate relation to the social world which is necessary for learning racism . . . . This is revealed most clearly in the failure of robots to manage common-sense knowledge in its tacit and social forms—a problem that has come to be known as the common-sense knowledge problem.[88]

Clearly, algorithms will continue to make mistakes; humans make them too. Algorithmic decision-making should not be labeled as biased simply because such mistakes exist. If anything, through additional data and iteration, algorithms will learn to better handle the messiness and staggering volume of unstructured data.

5. Retail and Price Discrimination

In April 2016, Bloomberg reported broad disparities between the availability of Amazon Prime's Free Same-Day Delivery service in minority neighborhoods compared to white areas of several towns.[89] The lack of free delivery services from the vast online superstore compounded the fact that brick-and-mortar retailers already shunned black areas, forcing residents to travel farther, and sometimes pay more, to obtain household necessities. According to Bloomberg's analysis, the service area for Amazon same-day delivery in six major cities excluded predominantly black ZIP codes to varying degrees. In one city, ZIP codes encompassing primarily black neighborhoods were excluded from same-day service despite

---

[88] Matthew T. Nowachek, *Why Robots Can't Become Racist, and Why Humans Can*, PHAENEX, Spring/Summer 2014, at 57, 58–59.

[89] David Ingold & Spencer Soper, *Amazon Doesn't Consider the Race of Its Customers. Should It?*, BLOOMBERG (Apr. 21, 2016), http://www.bloomberg.com/graphics/2016-amazon-same-day.

the fact that neighborhoods that surrounded them on all sides were eligible.

In response to an outpour of public criticism and scorn from city and state leaders, Amazon pledged to expand its same-day delivery area to include all neighborhoods.[90] At the same time, the company denied that the ethnic composition of neighborhoods was a factor in drawing up delivery maps. It stated instead that rollout of its same-day delivery service focused on ZIP codes where there was a high concentration of Amazon Prime members as well as on logistical delivery concerns.

In deciding where to deploy same-day delivery, Amazon apparently implemented a policy-neutral algorithm, which selected locations based on practical considerations.[91] Should Amazon have been required to also weigh equitable considerations, providing equal access for different racial groups? Should brick-and-mortar stores be subject to the same standard? For example, should Whole Foods be forced to open stores even in areas where a big enough market does not exist?

Moreover, the zeitgeist about these social issues may be fickle and in constant flux. Just a few years ago, Amazon was perceived as a threat to bookstores and to cultural neighborhood hubs.[92] In those days, some of the same critics who now fight for same-day delivery held arms together in an attempt to protect small neighborhood businesses by keeping Amazon out. In contrast, today the political consensus appears to be clear, viewing equal access to superior Amazon delivery as a social good, or even a human right.

In many circumstances, squeezing profits out of retail markets may have disparate effects on different populations. In September 2015, *ProPublica* published a research project by a team of reporters, revealing that The Princeton Review's online SAT tutoring packages varied substantially in price depending on where

---

[90] *Id.*

[91] *Id.*

[92] Lynn Neary, *End of Days for Bookstores? Not If They Can Help It*, NPR (Dec. 14, 2010, 9:01 AM), http://www.npr.org/2010/12/14/132026420/end-of-days-for-bookstores-not-if-they-can-help-it.

customers lived.[93] Customers who entered certain ZIP codes into the company's website were offered The Princeton Review's Premier course for an amount equal to $6,600; in other ZIP codes, customers had to pay $8,400 for the same course. More troubling, the research demonstrated a strong correlation between a customer's racial background and the offered price. Asian Americans were two times more likely to be represented in expensively priced ZIP codes than non-Asians. The reporters dispelled the possibility that the disparate pricing reflected a correlation to household wealth.[94] Even in lower-income ZIP codes, Asian Americans had a disproportionate likelihood of being offered a high price.

Previous press reports highlighted other instances of similar price discrimination, or "weblining."[95] For example, the *Wall Street Journal* found that Staples displayed different prices on its website depending on the distance of a customer's estimated location from a rival brick-and-mortar store.[96] The report showed that areas with a higher average income tended to see discounted prices compared to higher prices seen by customers in poorer neighborhoods. The FTC noted, "[i]f such pricing results in consumers in poorer neighborhoods having to pay more for online products than consumers in affluent communities, where there is more competition from brick-and-mortar stores, these poorer communities would not realize the full competition benefit of online shopping."[97]

---

[93] Julia Angwin & Jeff Larson, *The Tiger Mom Tax: Asians Are Nearly Twice As Likely To Get a Higher Price from Princeton Review*, PROPUBLICA (Sept. 1, 2015, 10:00 AM), https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review.

[94] *Id.*

[95] *What Weblining Means for You and Your Online Privacy*, REPUTATION DEFENDER (Sept. 29, 2011), https://www.reputationdefender.com/blog/privacy/what-weblining-means-you-and-your-online-privacy.

[96] Jennifer Valentino-Devries, Jeremy Singer-Vine & Ashkan Soltani, *Websites Vary Prices, Deals Based on Users' Information*, WALL ST. J. (Dec. 24, 2012), https://www.wsj.com/articles/SB10001424127887323777204578189391813881534.

[97] FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? 11 (2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

In these cases, assuming a lack of malice or intent to discriminate, companies apparently used policy-neutral algorithms to advance their bottom line goals. It was not the algorithm that was biased or discriminatory; the outcome was really driven by supply and demand. For example, The Princeton Review's higher prices in Asian neighborhoods may reflect the higher marginal cost of commissioning additional tutors in an area of high demand. Should the company subsidize those areas in order to avoid the appearance of price discrimination? That question has little to do with algorithmic discrimination.

B. Tech-Mediated Human Decision-making

This section provides examples of cases characterized by technology-mediated human decision-making. In these cases, decision makers may use algorithms to inform decisions or facilitate access to resources, but the bias and discrimination are the doings of human actors and should not be attributed to the algorithms.

1. Urban Potholes

In 2013, Boston adopted an innovative solution to combat the common municipal problem of road potholes. The city introduced "Street Bump," an app using the motion-sensing capabilities of smartphones to automatically report information to municipal government about the condition of the streets users drive on. When a user's car hit a pothole, their phone recorded the shock and sent it to a data hub, which combined the information from many other phones to pinpoint problem areas on streets to be repaired.[98] Interestingly, "Street Bump" reported more potholes in wealthy areas of the city than in poor ones. In retrospect, analysts discovered that the result, which could have regressively diverted urban resources from the poor to the rich, was driven by the unequal distribution of smart phones and app usage across the population. Affluent neighborhoods had more smart-phone and app users than poorer ones, causing the discrepancy.[99]

---

[98] *See* BOSTON MAYOR'S OFFICE OF NEW URBAN MECHS., *About Street Bump*, http://www.streetbump.org/about (last visited Sept. 21, 2017).

[99] Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Despite being presented as evidence for the risks of algorithmic decision-making, the Boston Street Bump app had little to do with data-driven discrimination. If the app were programmed to apportion greater weight to reports coming from wealthier neighborhoods than poorer ones, for example, critics could rightly blame it for class-based discrimination. But that was not the case with Street Bump, which simply created a seamless way to report and help fix a common urban flaw. In this case, where a higher density of smartphone users in wealthier neighborhoods created the concentration of reports, critics were not really faulting the app but rather the city's socio-economic fabric. Like many large American cities, Boston has racial, ethnic, and socio-economic fault lines, which transcend ownership and use of smartphones and apps.

What do these urban inequalities imply for services like Street Bump? Should cities avoid deploying new apps just because they help part, but not all, of their population? And against which backdrop should municipal leaders assess Street Bump's disparate impact? Perhaps the previous pothole reporting system—mailing complaints through the post or calling them in on the phone—was unbalanced as well? More generally, in an unequal society, every time an institution acts to improve a system, improving life for some citizens, it can be criticized for increasing—or at least not diminishing—existing disparities with persons who are worse off. Does that imply that until all disparities are purged urban systems should not improve?

In its Big Data Report, the White House credited Boston with discovering the biased reporting structure in time to prevent unjust resource allocation.[100] Using the algorithm to gather data was useful as long as it did not drive final policy decisions. "It took foresight to prevent an unequal outcome, and the results were worth it. The Street Bump app has to date recorded 36,992 'bumps,' helping

[100] EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014), at 51-52, https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

Boston identify road castings like manholes and utility covers, not potholes, as the biggest obstacle for drivers," the Report found.[101]

2. Sharing Economy

The sharing economy has upended the economic foundation of markets ranging from transportation and travel to finance and agriculture.[102] In online intermediated marketplaces, a balance must be struck between removing the identifying features of transacting parties to simplify and facilitate arms-length transactions and reducing users' anonymity to enhance trust and accountability. Alas, as soon as identifying features emerge, human biases creep in.

Harvard Business School professors Benjamin Edelman and Michael Luca investigated the extent of racial discrimination with hosts on Airbnb.[103] They demonstrated that holding location, rental characteristics, and quality constant, nonblack hosts were able to charge approximately 12 percent more than black hosts on the website. The authors concluded that "these differences highlight the risk of discrimination in online marketplaces."[104] Similar results

---

[101] *Id*. at 52.

[102] The sharing economy is a term used to describe a new way of organizing economic activity as a peer-to-peer commercial exchange, supplanting the traditional corporate-centered model. Some sharing economy platforms, such as Uber, Lyft, and Airbnb have ballooned in value to become some of the largest companies in the world. In July 2017, Uber, a private company, was worth an estimated $70 billion. *See* FED. TRADE COMM'N, THE "SHARING" ECONOMY: ISSUES FACING PLATFORMS, PARTICIPANTS & REGULATORS (2016), https://www.ftc.gov/system/files/documents/reports/sharing-economy-issues-facing-platforms-participants-regulators-federal-trade-commission-staff/p151200_ftc_staff_report_on_the_sharing_economy.pdf; ARUN SUNDARARAJAN, THE SHARING ECONOMY: THE END OF EMPLOYMENT AND THE RISE OF CROWD-BASED CAPITALISM (2016); *The Rise of the Sharing Economy*, ECONOMIST (Mar. 9, 2013), https://www.economist.com/news/leaders/21573104-internet-everything-hire-rise-sharing-economy.

[103] Airbnb is a digital platform for short term vacation rentals connecting between property owners and tourists all over the world. *See* About Us, AIRBNB, https://www.airbnb.com/ (last visited Sept. 21, 2017).

[104] Benjamin G. Edelman & Michael Luca, *Digital Discrimination: The Case of Airbnb.com* 4 (Harv. Bus. Sch., Working Paper No. 14–054, 2014), http://www.hbs.edu/faculty/Publication%20Files/Airbnb_92dd6086-6e46-4eaf-9cea-60fe5ba3c596.pdf.

have been reported by a group of researchers from San Francisco State University with respect to Hispanic and Asian hosts, who on average had a 9.6 percent and 9.3 percent lower list price on Airbnb relative to white counterparts, controlling for neighborhood property values, user reviews, and rental unit characteristics.[105]

Assessing potential supply-side discrimination, Edelman and Luca, together with Dan Svirsky, found that *Airbnb* renters with black-sounding names such as Tamika, Darnell, or Rasheed were 16 percent less likely to have their applications accepted than applicants with otherwise identical profiles and white-sounding names, such as Kristen or Brad.[106] The authors concluded that discrimination not only persists but also may be exacerbated, in online communities. They suggested that to reduce discrimination, Airbnb should conceal guest names, "just as it already prevents transmission of email addresses and phone numbers so that guests and hosts cannot circumvent Airbnb's platform and its fees."[107]

While Airbnb may apply various strategies to combat user bigotry, none of these cases demonstrate algorithmic discrimination. Instead of proving technology is biased, the reports demonstrate users are.

III. TAMING THE GOLEM

Data analysis allows for granular distinctions to be made between individual characteristics, traits, preferences, proclivities, and behaviors. Today, these capabilities are ubiquitously deployed to vet job applications, manage college admissions, drive predictive

---

[105] Venoo Kakar, Julisa Franco, Joel Voelz & Julia Wu, *Effects of Host Race Information on Airbnb Listing Prices in San Francisco* 9 (S.F. State U., MPRA Paper No. 69974, 2016), https://mpra.ub.uni-muenchen.de/69974/1/MPRA_paper_69974.pdf.

[106] Benjamin Edelman, Michael Luca & Dan Svirsky, *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment*, AM. ECON. J.: APP. ECON., Apr. 2017, at 1, 12–13, http://pubs.aeaweb.org/doi/pdfplus/10.1257/app.20160213. For a legal analysis, see generally Michael Todisco, *Share and Share Alike? Considering Racial Discrimination in the Nascent Room-sharing Economy,* 67 STAN. L. REV. ONLINE 121 (2015).

[107] Edelman et al., *supra* note 106, at 18.

policing, control government entitlement programs, and more. In making such impactful decisions, potential inaccuracies and biases could result in policy decisions that negatively impact the interests of minority, low-income, and underserved populations. This chapter sets forth issues to be addressed by policymakers seeking to implement a coherent approach to algorithmic decision-making.

A. Are Humans Better?

Traditionally, a common legal impulse in response to discrimination by automated decision-making is to require the involvement of a human operator at certain focal points.[108] However, human intervention could conceivably heighten the risk of manipulation and bias, further aggravating inaccuracies and discrimination risks. Indeed, historically, mechanized risk-based profiles were initially introduced in the mortgage industry as a *response* to the unequal treatment of loan officers toward borrowers.[109] Human decision makers were never immune to prejudice or bias. Concerns over the opaqueness of "black box" algorithmic decisions will not necessarily be resolved—and may even be amplified—by shifting discretion to opaque, undisciplined human decision-making.[110] After all, the ultimate "black box" is the one we have in our heads.

Kroll et al. state, "[t]he implicit (or explicit) biases of human decisionmakers can be difficult to find and root out, but we can peer

---

[108] *See, e.g.*, Council Directive 95/46, art. 15, 1995 O.J. (L 281) 32 (EC), http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1995:281:0031:0050:EN:PDF (directing member states to grant the right not to be subject to a decision of legal effect solely based on automated processing of data meant to evaluate personal aspects of that person, such as job performance or creditworthiness).

[109] *See* U.S. DEP'T OF JUSTICE, ATTORNEY GENERAL'S 2001 ANNUAL REPORT TO CONGRESS PURSUANT TO THE EQUAL CREDIT OPPORTUNITY ACT AMENDMENTS OF 1976 (2001), https://www.justice.gov/crt/housing-and-civil-enforcement-cases-documents-102.

[110] Ed Felten, *Algorithms Can Be More Accountable Than People*, CENTER FOR INFORMATION TECHNOLOGY POLICY, (Mar. 19, 2014), https://freedom-to-tinker.com/blog/felten/algorithms-can-be-more-accountable-than-people/.

into the 'brain' of an algorithm: computational processes and purpose specifications can be declared prior to use and verified afterward."[111] The same cannot be said, of course, about human—and particularly individual—decision-making processes.

For example, researchers recently examined why the public school system in Broward County, Florida, which has a large proportion of minority students, consistently identifies an overwhelming majority of white or Asian students as "gifted," leaving black and Hispanic students behind.[112] Black third graders in the district were only half as likely as whites to be included in programs for the gifted, and the deficit was nearly as large for Hispanics.

The researchers found that the district, which had relied on teachers and parents to refer children to classes for the gifted, was able to eliminate the gap by administering a universal standardized screening test. Their report concluded:

> Why did the new screening system find so many more gifted children, especially among blacks and Hispanics? It did not rely on teachers and parents to winnow students. The researchers found that teachers and parents were less likely to refer high-ability blacks and Hispanics, as well as children learning English as a second language, for I.Q. testing. The universal test leveled the playing field.[113]

In Broward County, the Golem yielded more unbiased, equitable results than human actors. The same could be true elsewhere. Before ceding to the impulse of introducing human decision-making at every juncture to fix algorithmic bias, policymakers should consider whether human analysts would in fact reduce or perhaps accentuate discrimination concerns.

B. Benchmarking Against the Status Quo

Instead of drawing comparisons to a utopian ideal, critics should contrast algorithmic decision-making with real-world processes, which could resemble the Broward County school administration, or an old-boy-network assessing candidates for employment or

---

[111] Kroll et al., *supra* note 2, at 634.

[112] Susan Dynarski, *Why Talented Black and Hispanic Students Can Go Undiscovered*, N.Y. TIMES: UPSHOT (Apr. 8, 2016), http://nyti.ms/1XvWN9V.

[113] *Id*.

applicants for higher education. Blaming the algorithm for existing social inequality and requiring back office tweaking to cleanse its results is myopic. Instead of solving societal problems, it would bury them under a gloss of political correctness and sanitized accounting.

As a recent White House report shows, data-driven decision-making can generate tremendous benefits, including social gains and enhanced fairness and opportunities.[114] In many cases, automated decisions, while imperfect, present a vast improvement against traditional human bias. In area after area, bringing human decision makers into the mix introduces or amplifies bias.

In *Big Data: A Tool for Fighting Discrimination and Empowering Groups*, the Future of Privacy Forum and the Anti-Defamation League reported a series of case studies demonstrating how businesses, governments, and civil society organizations leveraged data analytics to protect and empower vulnerable groups, including by providing access to job markets, uncovering discriminatory practices, and creating new tools to improve education and assist those in need.[115] According to the White House report, data-driven decision-making can effectively reduce discrimination and promote fairness and opportunity, including expanding access to credit in low-income communities, removing subconscious human bias from hiring decisions and classrooms, and providing extra resources to at-risk students.[116]

To be sure, mistakes will be made, and algorithmic bias will remain. But to avoid throwing the baby out with the bathwater, critics should compare the consequences of algorithmic decisions to the prevailing status quo.

---

[114] EXEC. OFFICE OF THE PRESIDENT, ARTIFICIAL INTELLIGENCE, AUTOMATION, AND THE ECON. (2016), https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF.

[115] FUTURE OF PRIVACY F. & ANTI-DEFAMATION LEAGUE, BIG DATA: A TOOL FOR FIGHTING DISCRIMINATION AND EMPOWERING GROUPS (2014), https://fpf.org/wp-content/uploads/Big-Data-A-Tool-for-Fighting-Discrimination-and-Empowering-Groups-Report1.pdf.

[116] EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

C. Three Categories of Algorithmic Decision-making

This paper proposed two distinctions. First, it distinguished between processes weighed and decided by algorithms (e.g., Google search) and ones where algorithmic assessments are solicited but only as input for human decision-making (e.g., the Street Bump app). Second, it suggested assigning stricter transparency and ethical review obligations to companies using policy-directed algorithms than to ones drawing on policy-neutral algorithms. Importantly, this second distinction does not reflect a dichotomy, but rather a spectrum, since even policy neutral algorithms are designed and potentially skewed by human bias, and some editorial algorithms are subject to only cursory review to ensure readability, integrity, or appropriateness.

By definition and design, every algorithm discriminates; that is the very *purpose* of a methodology created to crunch through reams of personal data to tease out correlations and draw useful lessons and conclusions. At a basic level, all humans are much more alike than different. As a matter of fact, humans are incredibly similar to other primates, mammals, and other living things. As creatures that share more than seventy percent of their DNA with acorn worms, humans display precious little meaningful differences between themselves. Placing individuals into different categories is by definition discrimination, in the neutral sense of drawing distinctions between people and treating them differently.[117]

[117] Black's Law Dictionary defines "discrimination" as: "1) The effect of a law or established practice that confers privileges on a certain class or that denies privileges to a certain class because of race, age, sex, nationality, religion, or disability; 2) Differential treatment; esp., a failure to treat all persons equally when no reasonable distinction can be found between those favored and those not favored." *Discrimination*, BLACK'S LAW DICTIONARY (9th ed. 2009). "The word 'discrimination' carries a very different meaning in engineering conversations than it does in public policy. Among computer scientists, the word is a value-neutral synonym for differentiation or classification: a computer scientist might ask, for example, how often a facial recognition algorithm successfully discriminates between human faces and inanimate objects. But, for policymakers, 'discrimination' is most often a term of art for invidious, unacceptable distinctions among people—distinctions that either are, or reasonably might be, morally or legally prohibited." Kroll et al., *supra* note 2, at 678 n. 134.

To assess the ethical implications of discrimination, we need to unpack the meaning of the term, which is, of course, political and highly charged. As Robert Fullinwider wrote

> [m]any may be led to the false sense that they have actually made a moral argument by showing that the practice discriminates (distinguishes in favor of or against). The temptation is to move from "X distinguishes in favor of or against" to "X discriminates" to "X is wrong" without being aware of the equivocation involved.[118]

An ethical assessment of machine-driven distinctions requires a coherent theory of discrimination. The Golem cannot determine whether a distinction is ethical or not. In *Judged by the Tin Man: Individual Rights in the Age of Big Data*, we wrote, "[u]nless we come up with a comprehensive theory of discrimination that can be represented algorithmically, we have no rigorous way to distinguish between ethical and non-ethical machine-based discrimination."[119] We certainly should not expect the machine to make moral decisions that *we* have yet to make.

1. Illegal

When dealing with potential bias in algorithmic decision-making, it is useful to distinguish between three categories of cases. The first category includes cases in which discrimination is unlawful. Antidiscrimination laws typically govern decisions on credit, housing, and employment, and restrict the use of categories such as race, gender, disability, or age. Where redlining is considered illegal, so too should redlining through proxies and

---

[118] ROBERT K. FULLINWIDER, THE REVERSE DISCRIMINATION CONTROVERSY: A MORAL AND LEGAL ANALYSIS 12 (1980).

[119] Omer Tene & Jules Polonetsky, *Judged by the Tin Man: Individual Rights in the Age of Big Data*, 11 J. TELECOM. & HIGH TECH. L. 351, 356 (2013). There have been attempts of statistical testing for discrimination in big data analysis. *See, e.g.*, Salvatore Ruggieri, Dino Pedreschi & Franco Turini, *Data Mining for Discrimination Discovery*, 4 ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA, No. 2 Art. 9 (2010); Binh Thanh Luong, Salvatore Ruggieri & Franco Turini, *k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention*, *in* PROCEEDINGS OF THE 17TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 502 (2011). These efforts too must first coalesce around an agreed upon delineation of legitimate vs. illegitimate discrimination.

automated tools. The profound capability of computers to identify patterns in endless piles of unstructured data facilitates the masking of illegitimate discrimination behind mirrors and proxies.[120] Decision-making, automated or not, based on such criteria should be banned.

2. Shadow of the Law

The second category includes cases of discrimination in the shadow of the law, which may not be technically illegal, but are nevertheless clearly indefensible. For example, even if it is not illegal in a specific state for a private business to discriminate on the basis of sexual preference, corporate ethics and policies should not allow an organization to be associated with discrimination. Accordingly, industry leaders such as PayPal, Google, Apple, Facebook, and Charlotte-based Bank of America, the largest corporation in North Carolina, have recently cut back investment in that state because of its legislation discriminating against the rights of gay, lesbian, bisexual, and transgender people.[121]

3. Unregulated

The third category is more ambiguous and includes cases of price discrimination or ad targeting based on profiling different groups. Here, social values are fickle and unsettled. Have we decided why it is legitimate to market to pregnant women in one context (e.g., based on subscription to a magazine), but morally distasteful to do so in another (e.g., Target's compilation of a "pregnancy score" for shoppers)?[122] Can an employer ethically

---

[120] Kroll et al., *supra* note 2, at 682 ("A prejudiced decisionmaker could skew the training data or pick proxies for protected classes with the intent of generating discriminatory results.").

[121] Jonathan M. Katz & Erik Eckholm, *Anti-Gay Laws Bring Backlash in Mississippi and North Carolina*, N.Y. TIMES (Apr. 5, 2016), http://www.nytimes.com/2016/04/06/us/gay-rights-mississippi-north-carolina.html.

[122] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES: MAGAZINE, (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?mcubz=3 ("Target can buy data about your ethnicity, job history, the magazines you read, if you've ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of

decline to interview a job candidate because they see a picture of them drinking a beer on a social media site? Is it appropriate for a travel agency to price discriminate based on use of a computer operating system, which may possibly correlate to wealth? Should we preempt any form of discrimination by requiring companies to mail Porsche catalogs to everyone regardless of income? Should Victoria Secret or Pampers be required to target all shoppers regardless of gender or age? The social norms triggered by these cases remain unsettled. It is far from clear that companies should be required to socially engineer against such an uncertain ethical background.

Zeynep Tufekci writes, "[w]e use these algorithms to explore questions that have no right answer to begin with, so we don't even have a straightforward way to calibrate or correct them."[123] Some of our ethical and moral criteria are so fickle, nuanced, and culturally dependent that it is doubtful that an automated process will ever be capable of appropriately weighing them.

As Kroll et al. explain, "[t]echnical tools offer ways to ameliorate these problems, but they generally require a well-defined notion of what sort of fairness they are supposed to be enforcing."[124] Indeed, it is far from clear that we would even *want* a machine to obtain the ability to distinguish right from wrong.[125] Such a Golem—a "technological singularity"—could possibly cause more moral angst than a current dumbed-down version.[126]

With companies becoming the gateways to information in the digital economy—Google enabling access to knowledge and commerce, Facebook and Twitter to news and social connections,

---

coffee, paper towels, cereal or apple sauce, your political leanings, reading habits, charitable giving and the number of cars you own.").

[123] Zeynep Tufekci, *The Real Bias Built In at Facebook*, N.Y. TIMES (May 19, 2016), http://www.nytimes.com/2016/05/19/opinion/the-real-bias-built-in-at-facebook.html.

[124] Kroll et al., *supra* note 2, at 678.

[125] *See* Tene & Polonetsky, *supra* note 11.

[126] *See* Zeynep Tufekci, *The Year We Get Creeped Out by Algorithms*, NEIMAN LAB (Dec. 2014), http://www.niemanlab.org/2014/12/the-year-we-get-creeped-out-by-algorithms.

Uber to transportation, Yelp to restaurant reviews, and more—concerns have grown around algorithmic gerrymandering and digital content manipulation. Scholars have sought ways to make digital platforms that wield power over information flows more accountable to users.

Jack Balkin proposed treating digital platforms as information fiduciaries, the lawyers and doctors of a digital age.[127] He asked, "[s]hould we treat certain online businesses, because of their importance to people's lives, and the degree of trust and confidence that people inevitably must place in these businesses, in the same way that we treat certain professional and other fiduciary relationships?"[128] Woodrow Hartzog and Neil Richards agreed that fiduciary obligations and trust should govern data relationships between companies and individuals.[129]

Another solution is to incentivize users away from prejudice and bias through smart user interface design, without editing algorithmic results. For example, Nextdoor.com, "a free neighborhood bulletin board where locals trade tips about plumbers, gossip about new shops, and alert each other about break-ins," has recently altered its interface to curtail racist comments and conjecture.[130] While not editing results of organic searches for "beautiful babies" or "beautiful women," which yield results overwhelmingly dominated

---

[127] *See* Jack Balkin, *Information Fiduciaries in the Digital Age*, BALKINIZATION (Mar. 5, 2014), http://balkin.blogspot.com/2014/03/information-fiduciaries-in-digital-age.html; Jonathan Zittrain, *Facebook Could Decide an Election Without Anyone Ever Finding Out*, NEW REPUBLIC (June 1, 2014), https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering.

[128] Balkin, *supra* note 127; *see also* Balkin, *supra* note 38. In 2008, we offered a similar argument with respect to search engines: "The law of confidentiality solves the problem of trust between patient and physician, customer and banker, and other additional fiduciary relationships. It should be applied equally to the transaction between search engines and users." Omer Tene, *What Google Knows: Privacy and Internet Search Engines*, 2008 UTAH L. REV. 1433, 1490 (2008).

[129] Neil Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 19 STAN. TECH. L. REV. 431 (2016).

[130] Nanette Asimov, *Nextdoor Social Site Cracks Down on Fearmongering*, S.F. CHRON. (May 10, 2016), http://bit.ly/1rQEj7H.

by white individuals, Google adds a toolbar with buttons enabling a user to easily select images of "African American," "Hispanic," "Native American," "Arab," and other minorities. Thus, without actually editing the results of a policy-neutral algorithm, Google provides an opportunity for multicultural outcomes.

Kroll et al. suggest a line of technical tools—which still have to be made—intended to assure fidelity to substantive policy choices and enforce non-discrimination.[131] These include the emerging science of fair classification in machine learning, dubbed by Dwork et al. "fairness through awareness."[132] Moritz Hardt et al. built on this approach, creating a methodology for measuring and preventing discrimination using a set of sensitive attributes, which is based on the idea that individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome.[133] Another tool is differential privacy, which ensures that with respect to any classification, an observer cannot determine whether or not a given individual was a member of a protected group.[134] Also available is zero-knowledge proof, which is a cryptographic tool that allows a decision maker, as part of a cryptographic commitment, to prove that a decision or policy has a certain property—for example, that a group of people belong to a certain category or class—without having to reveal either how that property is known or what the decision policy is.[135]

---

[131] Kroll et al., *supra* note 2, at 637.

[132] CYNTHIA DWORK ET AL., FAIRNESS THROUGH AWARENESS 214 (2011), https://arxiv.org/pdf/1104.3913.pdf.

[133] MORITZ HARDT ET AL., EQUALITY OF OPPORTUNITY IN SUPERVISED LEARNING, (2016), https://arxiv.org/pdf/1610.02413.pdf; *see also* MATTHEW JOSEPH ET AL., FAIRNESS IN LEARNING: CLASSIC AND CONTEXTUAL BANDITS (2016), https://arxiv.org/pdf/1605.07139.pdf.

[134] CYNTHIA DWORK, DIFFERENTIAL PRIVACY 1 (2006), https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf.

[135] *See* Kroll et al., *supra* note 2, at 667; Oded Goldreich & Yair Oren, *Definitions and Properties of Zero-knowledge Proof Systems*, 7 J. CRYPTO. 1, 7 (1994).

In *Beyond IRBs: Ethical Guidelines for Data Research*,[136] we built on a proposal by Ryan Calo to establish corporate "Consumer Subject Review Boards" to address ethical questions about data research in the private sector.[137] Calo suggested that organizations should "take a page from biomedical and behavioral science" and create small committees with diverse expertise that could operate according to predetermined principles for ethical use of data. The idea resonated in the White House legislative initiative, the Consumer Privacy Bill of Rights Act of 2015, which requires the establishment of a Privacy Review Board to vet non-contextual data uses.[138] In Europe, the European Data Protection Supervisor has recently announced the creation of an advisory group to explore the relationships between human rights, technology, markets, and business models from an ethical perspective, with particular attention to the implications for the rights to privacy and data protection in the digital environment.[139] Such new institutions would set forth policies for the design, deployment, and review of algorithms; assess the appropriateness of human intervention; and hold both internal and external accountability obligations. These institutions are a good start in thinking how to better build internal processes for algorithmic transparency and accountability.

CONCLUSION

With the political process polarized and consensus building impaired, society increasingly outsources difficult policy choices to companies and their automated machines. The allegations that Facebook's editorial decisions reflect a political bias highlight the risk of requiring or even allowing companies to scrub algorithms of

---

[136] Omer Tene & Jules Polonetsky, *Beyond IRBs: Ethical Guidelines for Data Research*, 72 WASH. & LEE L. REV. ONLINE 458 (2016), http://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/7.

[137] Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 STAN. L. REV. ONLINE 97, 102 (2013), http://pacscenter.stanford.edu/wp-content/uploads/2016/05/Calo-Consumer-Subject-Review-Boards.pdf.

[138] WHITE HOUSE, ADMINISTRATION DISCUSSION DRAFT: CONSUMER PRIVACY BILL OF RIGHTS ACT OF 2015 § 103(c) (2015), https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf.

[139] *Ethics,* EUR. DATA PROT. SUPERVISOR (2015), https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/Ethics.

perceived social biases. Such decisions and normative choices are better made in broad sunlight rather than under the guise of corporate non-disclosure agreements.

In a deeply divided political climate where societal views on discrimination based on race, ethnicity, income, age, and gender remain polarized and in constant flux, simply shifting the conversation from town hall to engineering shop is unsatisfactory. Social critics who largely embrace liberal values should note that by requiring code to adjust to their views, they pave a path for similar digital manipulation by governments and corporations. This can lead to unsettling results considering the ability of such organizations to tailor content and ads, manipulate the visibility of friends' visible posts and social interactions, and manage offers in the commercial and political sphere, often without individuals' knowledge or understanding.[140] It can also corrupt data used by researchers to advance scientific knowledge, given the *ex post* difficulty to distinguish real-world results from intentionally injected algorithmic noise.

To be sure, digital platforms and algorithms should encode the law and widely accepted social values and norms. Credit institutions, for example, must not differentiate between borrowers based on race, regardless if the decision is automated or not. But relying on machine intelligence to prevent the surfacing of the fault lines that divide diverse communities risks a future of unaccountable corporate control and opaque social manipulation.

In the 2001 Hollywood hit *Shallow Hal*, the protagonist, who is fixated on the physical beauty of women he dates and does not see them for who they really are, is hypnotized into visualizing women's inner beauty instead.[141] But by projecting the inner self as an external trait, the process leaves Hal preoccupied with aesthetic qualities. Policymakers should beware of calling on companies to provide users with a hypnotized *Shallow Hal* view of the world. The goal should not be to cosmetically alter Hal's perception; it should be to

---

[140] Ira S. Rubinstein, *Voter Privacy in the Age of Big Data*, 2014 WISC. L. REV. 861, 914 (2014).

[141] *See Shallow Hal Plot*, IMDB (2001), http://www.imdb.com/title/tt0256380/plotsummary.

educate Hal so he learns not to judge people based on appearance, or—if necessary and in line with social consensus—legislate against appearance-based discrimination, so Hal can make better choices down the road.