PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

presents

Inaugural Annual Privacy Summit

Session 9

Data Ethics – Important Issues of the Day (Joint Session)

MCLE: 1.0 Hours

Friday, February 10, 2023
4:30 p.m. – 5:30 p.m.

Speakers:

Professor Mark McKenna, UCLA Law School
Michael Karanicolas, Executive Director, Institute for Technology, Law & Policy, UCLA
Jennie Wang VonCannon, Partner, Crowell & Moring LLP

Conference Reference Materials

PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

Data Ethics – Important Issues of the Day

Nick Ginger, Senior Counsel, City National Bank
Professor Mark McKenna, UCLA Law School
Michael Karanicolas, Executive Director, Institute for Technology, Law & Policy, UCLA
Jennie Wang VonCannon, Partner, Crowell & Moring LLP

PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

CALIFORNIA RULE OF PROFESSIONAL CONDUCT RULE 1.1(A): COMPETENCE (MAR. 22, 2021)

- "A lawyer shall not intentionally, recklessly, with gross negligence, or repeatedly fail to perform legal services with competence."
- Comment 1: "duty to keep abreast of the changes in the law and its practice, including the benefits and risks associated with relevant technology"

TECHNOLOGY COMPETENCE

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Be able to operate technology used in practice of law
- Stay abreast of advances and additional technologies
- Understand security risks and how to mitigate them

BENEFITS

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Efficiency
- Increased volume
- Cost savings
- Remote work

RISKS

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Data breaches
- Cyber attacks
- Phishing
- Ransomware

PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

DATA EXFILTRATION VS. RANSOMWARE

OTHER RULES IMPLICATED AND AFFECTED

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Rule 1.6 – Confidential Client Information (CCI)
  - Business & Professions Code § 6068(e)(1): duty of lawyer "to maintain inviolate the confidence, and at every peril to himself or herself to preserve the secrets, of his or her client."
  - Make reasonable efforts to prevent inadvertent or unauthorized disclosure of or unauthorized access to CCI

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

OTHER RULES IMPLICATED AND AFFECTED

Rules 5.1, 5.2, 5.3 – Supervision of Attorneys & Non-Attorneys

Policies, procedures, training re: use of reasonably secure methods of electronic communications with clients

Instruct, supervise re: reasonable measures for access to and storage of those communications

Ensure policies implemented and kept up-to-date

PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

OTHER RULES IMPLICATED AND AFFECTED

- Rule 1.4 comment 2 and Rule 1.16(e)(1): attorney may send by "electronic means" documents to client upon client's request or upon termination of representation
- Rule 4.4 and Rule 1.0.1(n): notification requirements for material inadvertently produced via email
- Rule 7.2(a): email, social media post can be "advertisement"
- Rule 7.5 and comment: logos and website domains can be "communication" that is "false and misleading"

PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

COMMITTEE ON PROFESSIONAL RESPONSIBILITY AND CONDUCT (COPRAC) FORMAL OPINIONS

- 2005-168: Websites and Confidentiality
- 2007-174: Electronic Version of Client Files
- 2010-179: Transmitting and Storing Confidential Information
- 2012-184: Virtual Law Office
- 2012-186: Social Networking
- 2013-188: Confidential Information and Unsolicited Emails
- 2015-193: ESI and Discovery Requests
- 2016-196: Attorney Blogging
- 2020-203: Data Breaches

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

COPRAC FORMAL OP. 2010-179: TRANSMITTING & STORING CONFIDENTIAL INFORMATION

- Issue: using technology that is susceptible to unauthorized access to transmit or store confidential client information (CCI)
- Scenario: attorney uses a work-issued laptop to do legal research for client at a coffee shop and uses its public wireless Internet connection

PRIVACY LAW

CALIFORNIA LAWYERS ASSOCIATION

COPRAC FORMAL OP. 2010-179: TRANSMITTING & STORING CONFIDENTIAL INFORMATION

- Whether attorney violates duties of confidentiality and competence depends on:
  - Level of security of the technology, including whether reasonable precautions can be taken to increase security
  - Legal ramifications to malicious actor
  - Degree of sensitivity of information
  - Possible impact on client of inadvertent disclosure
  - Urgency of situation
  - Client's instructions and circumstances

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

COPRAC FORMAL OP. 2010-179: TRANSMITTING & STORING CONFIDENTIAL INFORMATION

- Attorney risks violating duties of confidentiality and competence by using public wireless connection if she does not also take precautions (*i.e.*, encryption, hotspot, VPN)
- Avoid public WiFi
- Or notify client of risks and seek informed consent to use it

COPRAC FORMAL OP. 2020-203: DATA BREACHES

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Issue: attorney's ethical obligations re: unauthorized access of electronically stored confidential client information
- Four scenarios:
  A. Laptop stolen, immediately remotely locked down and wiped clean
  B. Smartphone with 4-character passcode left at restaurant recovered the next day with no indication of access
  C. Firm paid ransomware demand and regained access to data; no CCI accessed and no matters negatively impacted by delay
  D. Attorney logged onto fake network at coffee shop; malicious actor accessed files on laptop related to client's patents

COPRAC FORMAL OP. 2020-203: DATA BREACHES

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Duty of disclosure [Rule 1.4(a)(3) and Bus. Prof. Code § 6068(m)]: must keep clients "reasonably informed about significant developments"
    - Misappropriation, destruction, or compromise of CCI
- Data breach that significantly impairs lawyer's ability to provide legal services

COPRAC FORMAL OP. 2020-203: DATA BREACHES

PRIVACY LAW
CALIFORNIA LAWYERS ASSOCIATION

- Data breach response plan
  - Monitor for data breaches
  - When breach is suspected or detected, ”act reasonably and promptly to stop the breach and mitigate [resulting] damage.”
  - Investigate and determine what happened—
    - Which clients affected
    - Amount and sensitivity of CCI involved
    - Likelihood that information has been/will be misused
- Get help from an expert in cybersecurity and data privacy

AI and Bias[1]

Introduction

As a growing proportion of our lives are governed by AI systems in both the private and public sphere, questions related to their accuracy and fairness have become increasingly pressing. Concerns about bias may seem counter-intuitive, since proponents of AI often point to its ability to remove ordinary markers of human bias from decision-making, and replace subjective assessments around, say, a person's trustworthiness or neediness, with mechanically generated values.[2] However, there is a volume of research which demonstrates that not only can AI systems introduce novel harms and discriminatory impacts, but that biased or discriminatory algorithms may be even more dangerous than human decision-makers because algorithms hide behind a veneer of neutrality.

In this section, we introduce the origins of bias in automated decision-making as well as its impacts. We end by considering these impacts' engagement with key legal concepts, and the state of legal scholarship in assessing these questions. The chapter begins with an overview of whether and how the legal system already addresses bias.

Bias and the Legal System

To frame our understanding of bias in AI systems properly, it is important first to consider a few avenues by which our legal system engages with both implicit and explicit bias. On an individual level, both lawyers and judges are expected to avoid discriminatory or harassing conduct. The American Bar Association's *Model Rules of Professional Conduct* section on "Misconduct" prohibits "conduct that the lawyer knows or reasonably should know is harassment or discrimination on the basis of race, sex, religion, national origin, ethnicity, disability, age, sexual orientation, gender identity, marital status or socioeconomic status in conduct related to the practice of law."[3] Judges, for their part, are required to perform their duties without bias or prejudice, to refrain from manifesting bias or prejudice, and to "administer justice without respect to persons."[4]

Historically, the most common legal questions related to bias typically manifested around employment or housing discrimination.[5] More recently, the criminal justice system has been a

---

[1] By Michael Karanicolas, Executive Director, UCLA Institute for Technology, Law & Policy, and Mallory Knodel, Chief Technology Officer, Center for Democracy and Technology. Thanks to Alessia Zornetta for her research and contributions.

[2] *See, e.g.,* Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making*, 22 STAN. TECH. L. REV. 290 (2019).

[3] MODEL RULES OF PROF'L CONDUCT, R. 8.4 cmt. 3 (2020).

[4] MODEL CODE OF JUDICIAL CONDUCT R. 2.3 (2020); 28 U.S.C. § 453 (2006).

[5] *See, e.g.,* Brown v. Board of Education of Topeka, 347 U.S. 483 (1954); *a landmark case on school integration, and* Buchanan v. Warley, 245 U.S. 60 (1917), *which invalidated a city ordinance banning the sale of real property in particular neighborhoods to blacks.*

major area of focus, including discriminatory conduct by police,[6] lawyers,[7] judges,[8] jurors,[9] witnesses,[10] and even court personnel.[11]

These different categories may be further subdivided to include both **conscious and unconscious bias**, otherwise known as **explicit and implicit bias**, with the latter now being widely accepted as having a broad and significant impact across a range of decision-making and other cognitive functions.[12] Critically, the existence of implicit biases, even powerful ones, does not mean that individuals will always act in biased ways, particularly since these biases may be **consciously overridden**.[13] Nonetheless, as the legal profession has come to recognize the impact of bias on decision-making and outcomes, it has led to an imperative to consider the impacts of **structural biases[14]** rather than attempting to root out overtly prejudiced individuals.

While there is no unified doctrine which the legal system uses to address bias, there are a number of principles which are relevant towards considerations of bias. First and foremost, the Fourteenth Amendment to the U.S. Constitution prohibits state governments from denying a person within their jurisdiction the equal protection of its laws.[15] As a result of the Fifth Amendment, the same standards apply to the decisions of the federal government, such as prosecutorial decisions.[16] However, since 1976 the Supreme Court has required that plaintiffs show a discriminatory intent in order to establish a violation,[17] though this standard has been criticized for being outdated in line with the volume of evidence related to the impacts of implicit bias.[18] Nonetheless, a law or policy which is neutral on its face will not be invalid under the Equal Protection Clause by virtue of having a more pronounced impact on one protected group than another. Direct intent is rare among AI systems, which generally return biased or discriminatory outcomes as a reflection of data or design flaws as opposed to overt instructions. As discussed in more detail in the following section, human bias can be introduced at each stage of development of AI systems. Together, these

---

[6] Paul Butler, *Equal Protection and White Supremacy*, 112 NW. U. L. REV. 1457, 1461- 62 (2018).

[7] Irene Oritseweyinmi Joe, *Regulating Implicit Bias in the Federal Criminal Process*, 108 CALIFORNIA LAW REVIEW 965 970-974 (2020).

[8] Chris Guthrie, Jeffrey John Rachlinski, Sheri Lynn Johnson & Andrew J. Wistrich, *Does unconscious racial bias affect trial judges?,* 84 NOTRE DAME LAW REVIEW 1195 (2009).

[9] https://pubmed.ncbi.nlm.nih.gov/10508569/.

[10] John P. Rutledge, *They All Look Alike: The Inaccuracy of Cross-Racial Identifications*, 28 AM. J. CRIM. L. 207, 211-14 (2001).

[11] Debra Lyn Bassett, *Deconstruct and Superstruct: Examining Bias Across the Legal System*, 46 UC DAVIS LAW REVIEW 1563, 1579 (2013).

[12] *See, e.g.,* Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464 (1998), *which has been particularly influential in driving understandings of implicit bias.*

[13] Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CALIFORNIA LAW REVIEW 969, 974 (2006).

[14] Structural bias, otherwise known as "institutional" or "societal" bias, is a process by which individuals may be treated unfairly as a result of their (perceived) membership in a particular category, including via the application of ostensibly neutral rules, to the extent such rules may reflect historical institutional arrangements which produced asymmetric outcomes.

[15] U.S. CONST. amend. XIV, § 1.

[16] Boiling v. Sharpe, 347 U.S. 497, 499 (1954).

[17] Washington v. Davis*,* 426 U.S. 229 (1976).

[18] *See, e.g.,* Yvonne Elosiebo, *Implicit Bias and Equal Protection: A Paradigm Shift*, 42 N.Y.U. REVIEW OF LAW & SOCIAL CHANGE 451 (2018), *which proposes a standard of discriminatory negligence for Equal Protection violations.*

characteristics suggest that the Equal Protection Clause may not be a major source for developing case law in this space or, alternatively, that existing precedent is ill-suited to combat discrimination in an administrative context which is increasingly governed by AI.[19]

Where decisions emanate from administrative agencies, American law also requires there to be a "rational connection between facts and judgment."[20] Although this standard accords significant deference to reviewing agency actions, it is potentially relevant to instances of bias or error among AI systems insofar as these decisions may fail to fulfill an adequate standard of **explainability**[21] and transparency.[22]

A 2020 study revealed that nearly forty-five (45) percent of federal agencies have used either AI or machine learning for a range of functions, including enforcing regulatory mandates and adjudicating government benefits and privileges.[23] The nature of these systems makes them resistant to meaningful review of the rationale underlying particular decisions.[24] This suggests that American law would benefit from the development of new judicial standards to deal specifically with AI-based adjudications, and particularly with the unique challenges in developing robust due process protections in the context of relatively inscrutable outputs from an AI decision-maker.

Where AI decisions emanate from private sector agencies, such as banks, potential or current employers, biased decision-making could engage the Civil Rights Act,[25] the Americans with Disabilities Act (ADA),[26] and Section 503 of the Rehabilitation Act,[27] among others. However, Supreme Court precedent generally requires either "intent" or "motive" in discrimination for disparate treatment, which are difficult to ascribe in the context of an AI decisionmaker due to the fact that machines typically do not possess intentionality the way that humans do.[28]

Absent this "intent" or "motive," plaintiffs may still succeed by demonstrating that a practice disparately impacts a particular protected group.[29] If this disparate impact is sufficiently demonstrated, the burden shifts to the defendant as to whether the practice is "consistent with business necessity." If a practice is found to meet the standard of business necessity, the plaintiff can still prevail if they are able to demonstrate that a less discriminating but equally valid practice was available which the employer did not use.[30]

---

[19] *Ibid.*

[20] Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 56 (1983).

[21] Explainability in AI generally refers to techniques that help approximate how a model produces an output, in order to support a standard of due process that is roughly analogous to the safeguards against arbitrary and capricious decision-making.

[22] Aram A. Gavoor, *The Impending Judicial Regulation of Artificial Intelligence in the Administrative State*, 97 NOTRE DAME LAW REVIEW REFLECTION 180, 184 (2022).

[23] DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO- FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 6–7 (2020), https://www-cdn.law.stanford.edu/wp- content/uploads/2020/02/ACUS-AI-Report.pdf.

[24] Pension Benefit Guar. Corp. v. LTV Corp., 496 U.S. 633, 654 (1990).

[25] 42 U.S.C. § 2000e (2012).

[26] 42 U.S.C. § 12101 (2017).

[27] Rehabilitation Act of 1973, Pub. L. No. 93-112, 87 Stat. 355, 393 (codified as amended at 29 U.S.C. § 793).

[28] Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335 n.15 (1977).

[29] Dothard v. Rawlinson, 433 U.S. 321, 329 (1977).

[30] 42 U.S.C. § 2000e-2(k) (2012); Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975).

In carrying out an assessment of discrimination by a private sector entity, the lack of explainability underlying AI decisions can be a significant complicating factor since it is difficult to pin down the underlying "practice" creating the disparate impact. A general decision to incorporate AI systems into the decision-making processes would be difficult, by itself, to establish such improper motive or intent, since many relatively benign factors, such as efficiency, might motivate a company to incorporate AI into its decision-making. Once again, the naturally human-centric manner in which jurisprudence has evolved, with its reliance on intent and rationale, runs into challenges in attempting to port the same standards to an AI decision-maker.[31]

Finally, it is worth considering bias more structurally in the legal profession. Law is a distinctly human activity, which is grounded in relatively subjective assessments of concepts such as reasonableness, consent, or intent.[32] However, every individual involved in the administration of justice, from the policy-makers who draft the legislation, to lawyers and judges who interpret these concepts, approaches them through their own set of lived experiences, with all of the baggage that can generate. None of us is truly a blank slate. Although there are a number of strategies for how bias can be confronted and mitigated, the growing recognition of the prevalence of bias has also been a key driver for diversity in the legal profession, in order to improve representation of perspectives and understandings of the law.[33] There are many novel aspects to how we think about bias in the context of AI, but at their core, these challenges are a continuation of a broader challenge to develop responsive and representative legal structures that reflect the spectrum of experiences and perspectives of the public they are meant to serve.

The Origins of Bias in AI Systems

The term "AI systems" is comprised of several technical concepts. AI is automation that aims to approximate human capability. Innovation in AI is often driven by the enthusiasm for exponentially increasing speed and scale of tasks through automation. Modern techniques to achieve automation include machine learning, deep learning and active learning.

Machine learning is a form of artificial intelligence algorithm that improves itself based on training data. The system "learns from experience." The way the machine "learns" depends on the algorithmic make-up of the system. Deep learning and active learning are more advanced techniques in which a system "learns how to learn" with (deep learning) or without (active learning) predetermined data sets.

Machine learning systems are enormous statistical interference engines with the capacity to generate outputs from the analysis of large inputs of data. Importantly, the data dependent nature of machine learning technology forms the basis of both the potentials and the pitfalls of contemporary artificial intelligence. Rather than eradicating human bias formed by the social and

---

[31] *For a more thorough discussion of this challenge in the employment context, see* Charles A. Sullivan, *Employing AI*, 63 VILLANOVA LAW REVIEW 395 (2018).

[32] Debra Lyn Bassett, *Deconstruct and Superstruct: Examining Bias Across the Legal System*, 46 UC DAVIS LAW REVIEW 1563, 1564 (2013).

[33] Sonia Sotomayor, *Lecture: 'A Latina Judge's Voice'*, N.Y. TIMES (May 14, 2009), https://www.nytimes.com/2009/05/15/us/politics/15judge.text.html.

historical processes, such as racist, sexist, or ageist preconceptions, human bias leaks into AI technologies at every turn, exposing the very social tenets of what is commonly understood as purely technological or, rather, technocratic. Machine learning bias, or what we refer to in this section as AI bias, occurs, then, when such algorithms produce outputs that are systemically prejudiced or discriminatory due to the underlying assumptions throughout various stages of the machine learning process.

There are many ways in which bias can find its way into AI: the structure of the data fed into the system, as well as the architecture of the algorithm both have a valence for the biased outputs that such systems may generate. This is particularly problematic when such systems are employed to automate processes in social institutions, because if the bias in the system is not addressed, artificial intelligence technologies risk automating the inequalities inherent in our social systems.

The governance of AI is a question that we will return to when we look at AI deployment, a crucial and iterative final stage. A technology-centric approach to address the fairness, accountability and transparency issues in AI systems relies on a framework that breaks down the machine learning process into its constituent parts: design, development and deployment. We take these phases in turn as we uncover where bias originates in AI systems.[34]

Bias in AI Design

The structural and human biases present in society appear in the design of AI systems from the problem-solution generation stage and persist through the early-development stage. In her book, *Race Against Technology,* scholar Ruha Benjamin exposes in great detail the ways "human decisions comprise the data and shape the design of algorithms, now hidden by the promise of neutrality and with the power to unjustly discriminate at a much larger scale than biased individuals."[35] For example, the problem is not only that predictive policing technologies are racially discriminatory, but that historically racialized groups are heavily policed and that predictive policing is seen as a way to scale up and automate the tasks required by over-policing communities of color.[36] In this way, bias has been introduced into an AI system before even one line of code has been written or one data point has been collected.

Other examples where the design phase introduces bias include targeted advertising. Although advertising merely aims to sell products to consumers, targeting is made possible through AI systems that are designed to take in data about individuals and code them as "interests." Although targeted advertising may not explicitly aim to capture an individual's race, many targeted advertising systems nevertheless are able to code race as interests, through preferences for hair products, food, fashion or music to the degree that large advertising platforms promote their

[34] Vidushi Marda, '*Governance with Teeth*', ARTICLE 19 (April 2019), https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

[35] Ruha Benjamin, *Assessing Risk, Automating Racism: A health care algorithm reflects underlying racial bias in society,* 366 SCIENCE 6464, (2019).

[36] Where "racialized" refers to a socio-political process by which groups are ascribed a racial identity, whether or not members of the group self-identify as such; where "predictive policing" refers to technical tools and practices that use data and analytics to identify potential crimes; and where "over policing" refers to disproportionate police presence in marginalized communities that exacerbates poverty, marginalization and criminalization of individuals in those communities.

success at reaching audiences based on race. Other explicit captures of data to target by race include location and "ethnic affinity."

Another example that has a much greater potential for harm is when socio-economic status, including race, leads to inequalities in access to financial services, either because products are advertised differently based on perceived status or because the very determination of individuals' credit worthiness is made through the same data. It is clear to see how the design of targeted advertising of financial services might exacerbate the cycle of poverty, even if race and socio-economic status are not explicitly captured by advertising platforms.

Bias in AI Development

There exist structural, statistical, socio-technical and human bias in the data, training procedures and validation stages of AI development. Each of these is taken in turn, below, to expose the origins of bias in AI systems.

Collecting Data and Data Sets

Machine learning algorithms requires vast amounts of data on which it must learn. This data is a major driver of bias in AI. Some sources of data are explicitly biased, such as troves of photographs and notations originating in eugenics research.[37] But in all cases, because "most machine-learning tasks are trained on large, **annotated data sets**,[38] such methods of annotating training data can unintentionally produce data that encode gender, ethnic and cultural biases."[39] Although not all types of machine learning rely on predetermined data sets such as active learning, any potential model might therefore be built upon the broad **datification**[40] of our deeply biased world.

For the most common types of AI that are trained on, and learn from, pre-determined data sets or data sources, the characteristics of the datasets used in machine learning fundamentally influence an AI model's behavior. A model is unlikely to perform well when it encounters novel data if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases.

"As a first step, researchers — across a range of disciplines, government departments and industry — need to start investigating how differences in communities' access to information, wealth and basic services shape the data that AI systems train on."[41] Scholars Kate Crawford and Ryan Calo are referring to how the data sets used in AI systems might themselves reproduce existing stereotypes by categorizing and inputting already differential manners in which various social groups access public resources.

---

[37] https://magazine.jhsph.edu/2022/how-biased-data-and-algorithms-can-harm-health

[38] Sometimes referred to as "data labeling," annotating data sets involves adding tags or labels to data such that algorithms can learn to identify novel data as belonging to the same category.

[39] James Zou, Londa Schiebinger, *Design AI so that it's fair,* 559 NATURE, 324-326 (2018).

[40] Datification refers to the pervasive collection, generation, storage, and analysis of data that is driven by profit models that commodify data and data analysis in the form of unique predictions and insights.

[41] Kate Crawford, Ryan Calo, *There is a blind spot in AI research,* 538 NATURE, 311–313 (2016).

The way in which data is collected from people also has weight in shaping the data sets. Here, institutional guidelines, as well as policies in tandem with documentational records, should be considered, as both of these have a direct effect on the types of data gathered. For example, NIST maintains a database of mugshot photos, in addition to other standard reference data, that is open and available as a training data set, yet it is under-representative of the U.S. population in categories such as gender and race. Mitigating potential bias in the use of such sets requires an open reckoning with such context.

Training AI Models

Another key aspect of AI development takes place in the training of an AI model. The AI model in its most basic form defines how input is turned into output. Training an AI model to perform according to the problem-solution generation of the design phase means active human supervision of the machine learning process as it takes in data and presents outputs. Tweaking **parameters**,[42] **re-characterising data**,[43] using statistical methods, and adjusting features of the model itself are ways that AI models are trained.

Human bias can be introduced at this stage because humans are doing some degree of supervising along a spectrum. Even in "unsupervised" machine learning, human feedback about the quality of the output is used by the algorithm to learn and optimize for future performance. Whether or not an answer is "right" can itself be a source of bias: For instance, an AI system might be tasked with sorting job applications to ensure new employees "fit in" in a company that has a problem with monoculture.

Some fixes are discussed by Zou and Schiebinger: "Thus, technical care and social awareness must be brought to the building of data sets for training. Specifically, steps should be taken to ensure that such data sets are diverse and do not under-represent particular groups. This means going beyond convenient classifications — 'woman/man', 'black/white', and so on — which fail to capture the complexities of gender and ethnic identities."[44] The more complex the AI system, the harder it is to avoid bias, say, when images classified as 'nurse/doctor' are perfectly paired with 'woman/man'.[45]

Validating and Optimizing Outputs

As with the need to train or supervise machine learning, there must be an ongoing effort to validate outputs and optimize the model to increase accuracy, which some AI has been accused of lacking. Ensuing changes to the AI system can be made at any stage and take a variety of forms but are almost always determinations made by humans. Therefore, these determinations are points at which bias can be introduced. Adjusting the scope of the predetermined training data, data source, weights, parameters, and other changes in the data or how the AI will assess data can confirm or

[42] Parameters effectively comprise model behaviors, as they are static expressions of which characteristics are more important than others in a decision-making algorithmic system.
[43] Characterizing data: See "Annotating data sets".
[44] James Zou, Londa Schiebinger, *Design AI so that it's fair,* 559 NATURE, 324-326 (2018).
[45] https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

further entrench the biases introduced at the design and development stages. Such adjusting does, however, also offer points at which de-biasing can be done.

It has been suggested that – at the dataset level – "every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses."[46] This additional context and markers for datasets can increase transparency and accountability for developers. As the machine learning community at large has expressed a commitment to mitigate unwanted societal biases in machine learning models, it must work together to effectively reproduce machine learning results across diverse implementations. Another result of increasing data set transparency and sharing results is that researchers and practitioners can better select the appropriate datasets for their systems' goals.

At the training level, one technique to mitigate bias has been to statistically offset what is called "**word embeddings**"[47] like when nurse/doctor is equivalent to woman/man. Statistical offsets, e.g., consciously changing parameters to avoid work embeddings, leads to algorithms that "significantly reduce gender bias in embeddings while preserving the useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias."[48]

Others take the validation and optimization phase as an opportunity to introduce auditing mechanisms such as, "an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups," published by Joy Buolamwini and Timnit Gebru.[49]

Bias in AI Deployment

Once a trained machine learning model has been integrated into an AI system, its application to a problem-solution set requires mitigating the harms of the application of AI systems with fixes or improvements, or lessening its impacts. At the deployment stage, an AI system is often acting within an existing technocratic structure, for instance, to determine creditworthiness or the equitable delivery of social services. How an administrator or ultimate decision maker accounts for the AI system's output is a point at which bias can once again be introduced.

Virginia Eubanks' book, "Automating Inequality," dramatically shows how data collected for technological purposes becomes a means of reinforcing economic marginality, which she refers to as "collective red-flagging, a feedback loop of injustice" (Eubanks, 2018:7). She criticizes the

---

[46] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortmanvaughan, Hanna Wallach, Hal Daumé III, Crawford Kate, *Datasheets for Datasets,* PROCEEDINGS OF THE 5TH WORKSHOP ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING, Stockholm Sweden (2018).

[47] A term of art in natural language processing, "word embeddings" refers to when two or more encodings of the same string of text are taken by the algorithm to be similar to one another; for example: "nurse" is [taken to be] more similar to "woman" than it is to "doctor", which is more similar to "man" than "woman".

[48] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Sallgrama, Adam Kalai, *Manis to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,* 30TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2016), Barcelona Spain (2016).

[49] Buolamwini Joy, Gebru Timnit, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,* 81 PROCEEDINGS OF MACHINE LEARNING RESEARCH, 1–15 (2018).

notion that a model is less biased than a human caseworker, homeless service provider, or intake caller; "I find the philosophy that sees human beings as unknowable black boxes and machines as transparent deeply troubling."[50]

Eubanks poses two questions to assess the basic ethics of digital tools: (1) Does the tool increase the self-determination and agency of the poor? (2) Would the tool be tolerated if it was targeted at non-poor people?

These pointed questions can also be at odds with another set of questions about whether or not we "trust" the results of the AI system; whether or not the AI system's determination aligns with our own expectations. As Aylin Caliskan et al. write, "Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology."[51]

When models are built in one place, perhaps geographically, and then applied in another, unforeseen bias is much more likely to only be observable at the application phase. Shreya Shankar et al. write, "data sets appear to exhibit an observable amerocentric and eurocentric representation bias. Further, we analyze classifiers trained on these data sets to assess the impact of these training distributions and find strong differences in the relative performance on images from different locales. These results emphasize the need to ensure geo-representation when constructing data sets for use in the developing world."[52]

Yet there are additional questions beyond these pointed ethical considerations to more technological forms of AI governance such as the assessment of **fairness, accountability and transparency** (FAccT). The FAccT framework takes the middle ground between neutral and ethical technology to focus on questions of management: is it fair; is it accountable; is it transparent?[53]

Development of Tools to Assess "FAccT"

Engineers are often involved in designing the various auditing mechanisms that consider FAccT. R.K.E Bellamy et al. introduce an example of "a new open-source Python toolkit for algorithmic fairness, AI Fairness 360 (AIF360), released under an Apache v2.0 license (https://github.com/ibm/aif360). The main objectives of this toolkit are to help facilitate the

[50] ibid:168

[51] Caliskan Aylin, Bryson Joanna, Narayanan Arvind, *Semantics derived automatically from language corpora contain human-like biases,* 356 SCIENCE 6334, 183-186 (2017).

[52] Shankar Shreya, Halpern Yoni, Breck Eric, Atwood James, Wilson Jimbo, Sculley D, *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World,* Presented at NIPS 2017 WORKSHOP ON MACHINE LEARNING FOR THE DEVELOPING WORLD (2017).

[53] FAccT does not fix bias, it only helps as a framework for understanding it.

transition of fairness research algorithms for use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms."[54]

These mechanisms can be applied retroactively at every stage and need not wait until deployment. Sorelle Friedler et al. write, "we find that **fairness-preserving algorithms** tend to be sensitive to fluctuations in dataset composition (simulated in our benchmark by varying **training-test splits**) and to different forms of preprocessing, indicating that fairness interventions might be more brittle than previously thought."[55]

Yet there exists a final point along the spectrum from neutral, to FAccT, to ethical: Others have proposed a wider view of the social and human rights impacts of AI system deployment and application. Like Marda's work on AI governance cited above, Suresh Venkatsubramanian et al. write, "given that AI is no longer solely the domain of technologists but rather of society as a whole, we need tighter coupling of computer science and those disciplines that study society and societal values."[56] There is a now pervasive theory that altering the context in which AI engineering occurs, by innately considering human rights and society, will lead to more human rights aligned outcomes.

Yet the widest view acknowledges that indeed AI systems are inherently embedded in the human world, and the human world is biased. Thus, "Even with careful review of the algorithms and data sets, it may not be possible to delete all unwanted bias, particularly because AI systems learn from historical data, which encodes historical biases."[57]

Impacts of AI Bias on Law and Society

The expanding footprint of algorithms in our day to day lives, otherwise known as the **algorithmic turn**, has led to a growing body of scholarship related specifically to concerns about fairness and bias.[58] From our daily search for news and information, to our choice of romantic partners, to our ability to find a job or a home, or to access credit, our lives and decisions are increasingly governed by invisible formulas designed to deliver efficiency, profit, engagement, or any number of other

---

[54] R.K.E. Bellamy, et al., *AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*, 63 IBM JOURNAL OF RESEARCH AND DEVELOPMENT 4/5, 1-15 (2019).

[55] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth, *A comparative study of fairness-enhancing interventions in machine learning,* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAT* '19), 329–338 (2019).

[56] Suresh Venkatasubramanian, Nadya Bliss, Helen Nissbaum, Melanie Moses, *Interdisciplinary Approaches to Understanding Artificial Intelligence's Impact on Society* (2020). arXiv:2012.06057

[57] Drew Roselli, Jeanne Matthews, Nisha Talagala, *Managing Bias in AI,* WWW '19: COMPANION PROCEEDINGS OF THE 2019 WORLD WIDE WEB CONFERENCE, 539-544 (2019).

[58] Philip M. Napoli, On Automation in Media Industries: Integrating Algorithmic Media Production into Media Industries Scholarship, 1 MEDIA INDUSTRIES J. 33 (2014).

predetermined metrics.[59] The data intensive networks that underlie these decisions, and the cold and impassive way in which the results are delivered, creates an illusion of neutrality and fairness, especially as contrasted against the heavily subjective and instinct-driven processes that dominated most traditional decision-making.[60] However, as demonstrated by the previous section, it is all too common for these algorithmic systems to replicate, obfuscate, and entrench historical discriminatory structures, painting them with a veneer of objectivity and fairness while replicating many of their worst aspects.

AI Bias and Traditional Legal Notions of Discrimination[61]

An early focus of academics and civil society researchers has been around mapping these impacts, and their consequences, from a legal and social perspective. The traditional legal focus on **discrimination** related to employment has led to particular attention being devoted to the growing use of algorithms to sort and rank potential job applicants.[62] The use of AI in hiring decisions, even if only for triaging potential candidates, poses a threat not only because it can produce discriminatory results, but because it often does so through a facially non-discriminatory decision-making pattern. For example, one algorithm, which was designed to assess potential candidates based on the performance of existing employees, concluded that the two factors which most strongly correlated to strong performance were whether the candidate had played high school lacrosse, and whether their name was Jared.[63] Although neither categorization is discriminatory on a protected ground *per se*, these kinds of results are obviously going to be strongly associated with protected variables.

Even where an algorithm is specifically prohibited from decision-making based on protected variables, it may nonetheless cultivate stand-in variables as proxies to achieve the same, discriminatory result.[64] As noted in the second section, it is difficult for an algorithm to possess the discriminatory intent that is often required in order to make a legal challenge stick, making it difficult to develop robust structures for legal accountability. Moreover, the prevalence of these

---

[59] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

[60] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671, 1688 (2020).

[61] Bias and discrimination are not interchangeable terms: There can be all kinds of biases which, while problematic, should not give rise to a discrimination complaint. For instance: If some quirk in an AI system resulted in a bank hiring or elevating candidates with an even number of letters in their names, over persons with names with an odd number of letters, that would be a problem for the bank and the AI's designers – this would not necessarily be something that would be a basis for a lawsuit because it does not relate to a protected ground or category of persons. Discrimination is a legal standard. The question in legal proceedings will be whether bias has given rise to discrimination.

[62] Goldman Sachs, for example, announced in 2016 that it would rely on algorithmic models to automate not only their hiring decisions, but virtually all employee management decisions: Rob Copeland & Bradley Hope, *The World's Largest Hedge Fund Is Building an Algorithmic Model from Its Employees' Brains*, WALL ST. J. (Dec. 22, 2016), https://www.wsj.com/ articles/the-worlds-largest-hedge-fund-is-building-an-algorithmic-model-of-its-founders-brain-1482423694.

[63] See Dave Gershgorn, *Companies Are on the Hook if Their Hiring Algorithms Are Biased*, QUARTZ (Oct. 22, 2018), https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring- algorithms-are-biased.

[64] Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Aaron Rieke & Alan Mislove, *Algorithms that "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences*, https://mislove.org/publications/Lookalike-AIES.pdf.

proxy characteristics, which may be closely associated with, but distinct from, protected grounds, has the potential to allow decision-makers with prejudicial values to hide deliberate discrimination behind a mask of deniability.[65]

The use of AI in hiring also has the potential to metastasize the impact of these biases at scale. One of the primary benefits of technologies like AI decision-making is its speed and scalability, performing tasks that in an earlier age would have required an army of dedicated staff, and a significant investment of time. Although there is no question that discrimination in traditional human decision-making systems could also be structurally enabled and enforced, AI imposes an unprecedented level of uniformity and consistency to these decisions, calibrating them to a single standard. If this standard is biased or discriminatory, it can infect entire industries, causing far more harm than a single racist or sexist hiring manager. Although an AI decision-maker may be easier to retrain than a human, this requires that one first be able to isolate and diagnose the problem, which is challenging given the complex and opaque way that AI decisions are made.[66]

In considering appropriate policy and regulatory responses to these challenges, scholars of race and technology, such as Safiya Noble, Ruha Benjamin, and Ifeoma Ajunwa, have been at the leading edge of academic thinking around AI and bias, cautioning that without early intervention, the rollout of AI systems across the public and private sectors poses a grave threat to efforts to combat structural inequality and racism.[67]

One prominent theme has been that the challenges posed by biased AI decision-making need to be understood as more than mere technical glitches, which may be resolved through better code, better auditing, or a more judicious selection of training data. Ifeoma Ajunwa, in particular, has argued that the categorization of these challenges as technical problems is fundamentally misguided because there is always a human behind the curtain.[68] Moreover, she argues biased or discriminatory outcomes, even where directly delivered by an algorithm, should be viewed as a legal problem caused by anachronistic approaches towards regulating discriminatory decision-making, such as an overly deferential attitude towards employer choices.[69]

In other words, the spread of algorithms can not only exacerbate and reflect historical biases, but it can also create new opportunities for historical legal deficiencies to be exploited towards discriminatory ends. Solutions that aim to combat discrimination and bias should therefore not only target problems with the algorithms and their underlying data but should also aim to rectify these deficiencies in the surrounding legal or policy structure, such as through granting less deference to employers' decision-making which produces discriminatory outcomes.

---

[65] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 692-93 (2016).

[66] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671, 1679 (2020).

[67] *See, e.g.,* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Ruha Benjamin, *Assessing risk, automating racism* 366 SCIENCE 421 (2019); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671 (2020).

[68] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671, 1708 (2020).

[69] *Ibid.*

In a similar vein, Sandra G. Mayson has argued that the challenge from AI technologies is fundamental to their predictive nature, since any predictive system will, by its nature, project the inequities of the past onto the future.[70] On the use of AI in the criminal justice system, Professor Sandra Mayson argues that the problem is not just with the design of these systems, but with the very notion of predictive policing. Her response would therefore be to reconsider how we assess and respond to risk, since these interventions, and how we criminalize certain behaviors and respond to the emergence of criminal patterns as we have defined them, lies at the core of the discriminatory outputs that the criminal justice system generates. A similar point could be made regarding algorithms which screen prisoners' suitability for pretrial release or for parole, both of which will presumably reflect institutionally racist definitions of what constituted problematic behavior. A black parolee, for example, who was subject to a comparatively stricter level of police surveillance than a white counterpart, would naturally be more likely to be found in violation of their terms of release, ultimately generating a differential metric for the relative riskiness of these two groups.

Other legal scholarship has focused more specifically on technical fixes, or at the very least on reconsidering our approach to how AI is developed, implemented, and audited. Anupam Chander, writing in 2017, argued for the institution of a form of "algorithmic affirmative action," which would force a consideration of the disparate impacts of data and design related to categories where discrimination is legally prohibited (i.e., race, age, sex, religion, etc.[71]), and attempt to rectify these impacts through changes to the data or design which return less discriminatory results.[72]

A major challenge with implementing such technical solutions is the lack of public access to accurate information about how these systems were trained and are functioning. In one particularly well known case, an algorithmic recommendation tool meant to guide sentencing, known as COMPAS, was found to be returning results that were biased against black subjects, flagging them as a significantly greater risk to reoffend.[73] This case is particularly noteworthy in that there was an auditing procedure in place, which found that the system was fair because its overall accuracy rate in terms of predicting recidivism was roughly equivalent between the two racial groups. The audit neglected to consider that where the system failed it did so by placing black defendants in a riskier category, and white defendants in a less risky category. Legal scholars have suggested a range of tools aimed at mitigating this specific challenge, including developing and enforcing codes of conduct for the design of AI systems, and enhancing whistleblower protection rules to ensure that internal knowledge about discriminatory systems makes its way into the public realm.[74]

While challenges related to discrimination and structural bias are by no means a recent phenomenon, and certainly are not unique to AI, the salience of these technologies to the human

---

[70] Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L. J. 2218 (2019).

[71] This refers to certain categories ("protected classes") against which it is illegal to discriminate (again – race, religion, sex…), but the list may vary from statute to statute (and is a matter of some controversy and debate).

[72] Anupam Chander, *The Racist Algorithm?,* 115 MICH. L. REV. 1023 (2017).

[73] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[74] Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54 (2019).

condition, and their ability to both foreground new challenges and problematize existing and emerging social and legal challenges, mandates serious consideration in how judicial structures should approach challenges that will inevitably flow from the implementation of algorithms across the public and private sectors.[75] **Judges should expect that problems related to bias are likely to manifest, in one form or another, across virtually every field where AI decision-making has and will soon become popularized.** Although every case is unique, and some complaints will bear more merit than others within the context of the prevailing legal framework, judges should keep an open mind towards thinking through which approaches to the law may need to be adapted or reconsidered in light of the transformative impact of these technologies on human decision-making.

[75] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1 (2004).

To Err is Human, to Audit Divine: A Critical Assessment of Canada's AI Directive

By Michael Karanicolas*

1. Introduction

In recent years, an increasing number of global governments have been quietly incorporating automated decision-making systems into their governance processes. But while the potential efficiency gains from these systems are easy to see, their broader impact on core government functions is much less clear, not least because of a broader lack of transparency in how they are being rolled out, and a dearth of public discussion on how they should be used responsibly. From this perspective, Canada's passage, in April 2019, of the government's first *Directive on Automated Decision-Making* (the *Directive*),[1] is a welcome development, insofar as it kickstarts an important public conversation about how the federal government should use automated decision-making systems, and what safeguards should apply to their implementation. This Article will carry out an analysis of the strengths and weaknesses of the *Directive*, in the context of broader challenges around the use of automated decision-making systems, and provide substantive recommendations for improvement.

There is a large, and constantly growing, volume of academic research on inherent biases in our systems of governance and decision-making.[2] From well-trod discussions of discrimination related to race and gender, to a famous study that suggested judges' rulings varied substantially depending on how recently they had eaten,[3] substantial evidence supports the assertion that our ideal of a neutral adjudicator rendering decisions based purely on the facts in front of them is more a romantic fiction than a reflection of reality. However, while bias may be impossible to eliminate, the fundamental inevitability of it has allowed our legal system to develop various checks and balances aimed at mitigating its effects, such as the disclosure of potential conflicts, standardized legal tests aimed at corralling the decision-making process, requirements to explain (or justify) decisions, and various appeal mechanisms which aim at assessing whether a decision was reasonably concluded.

Given the fact that bias is such an inherently human flaw, it is somewhat paradoxical that the introduction of automated decision-making algorithms has generated such a robust debate over their tendency to return biased or discriminatory conclusions. However, it is precisely the humanity

---

* Wikimedia Fellow, Information Society Project, Yale Law School. I owe a debt of gratitude to Dr. Lisa Austin, whose syllabus and teaching at the University of Toronto were foundational for shaping the ideas expressed here, and to Benoit Deshaies who provided helpful feedback in finalizing this draft.

[1] Treasury Board of Canada Secretariat, *Directive on Automated Decision-Making* (2019), online: <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html>.

[2] See, for example, Robert R. Kuehn, "Bias in environmental agency decision making" (2015) 45:4 Environmental Law 957; Robert K. Christensen et al, "Race and Gender Bias in Three Administrative Contexts: Impact on Work Assignments in State Supreme Courts" (2012) 22:4 Journal of Public Administration Research and Theory 625; Ryan Bubb & Patrick L. Warren, "Optimal Agency Bias and Regulatory Review" (2014) 43:1 The Journal of Legal Studies 95.

[3] Shai Danziger et al, "Extraneous factors in judicial decisions" (2011) 108:17 Proceedings of the National Academy of Sciences 6889.

of this trait which makes algorithmic biases so challenging to deal with. Unlike a human decision-maker, which is likely to wear their biases on their sleeve, algorithms have the potential to be wolves in sheep's clothing, posing as neutral and purely mathematical arbiters, while returning results which may be tainted by underlying biases. Consequently, while the problem of bias in decision-makers is not itself a new phenomenon, and there is no question that our legal and regulatory systems still have a long way to go in terms of combating traditional forms of discrimination, the rise of algorithmic decision-making has generated new and unprecedented challenges to guaranteeing fair due process from governmental decision-making structures, since algorithmic biases can be more difficult to identify and mitigate.

2. Understanding the Challenge

Civil society and academic observers have raised a number of red flags with regard to the introduction and expanding use of automated decision-making algorithms. One obvious starting point to the discussion is a lack of transparency around these systems, since a robust conversation about the impact of algorithmic decision-making must begin with a proper accounting of the current state of implementation. There have been a number of individual reports of algorithmic technologies being employed in various governmental functions but, as of yet, no consolidated assessment of the full extent of their use.[4]

Although a lack of transparency is often the first concern that manifests in conversations about the use of algorithmic decision-making systems, transparency alone is not a panacea. Rather, the lack of transparency is a gatekeeping problem to addressing other fundamental issues of accountability, procedural fairness, and the potential for bias in algorithmic decisions. Ultimately, the very novelty of these systems presents a major challenge in developing an effective governance framework for their use, since this makes them resistant to the sorts of safeguards that we might apply to traditional decision-making processes. For example, algorithmic systems are typically unable to furnish an "explanation" for their decisions, the way a human decision-maker might.

The complexity of automated decision-making systems, whose operation relies not just on sophisticated code but also, in many cases, on an analysis of vast datasets, can also make potential problems more difficult to spot. These can be unwittingly ingrained into the system as a result of underlying biases in the algorithm (such as its incentive structure), its designers, or the data sets which it was trained on (which may be reflective of previous discriminatory policies). If there is an underlying bias in any of these areas, the algorithm may reinforce or magnify this in its outputs. Indeed, even if none of these components are themselves biased, they may interact together in a way which produces biased results. To borrow a phrase, the core of the challenge lies with the "unknown unknowns", problems whose scope or character are difficult to predict, may only manifest once a system is operational, and might only be discoverable through a careful, post-hoc study of the system and the results which it generated.

The challenge in spotting flaws in the results returned by an algorithmic system is well illustrated by the COMPAS case, where an algorithm which calculated risk levels for criminal

[4] See, for example, Zosia Bielski, "Toronto human rights lawyer sounds the alarm on Canada's plans to use AI in immigration", *Globe and Mail* (4 November 2018), online: <https://www.theglobeandmail.com/canada/article-toronto-human-rights-lawyer-sounds-the-alarm-on-canadas-plans-to-use/>.

defendants in a number of U.S. states was returning results that were biased against black subjects.[5] This case is particularly noteworthy in that there was an auditing procedure in place, and it found that the system was fair because its overall accuracy rate in terms of predicting recidivism was roughly equivalent between the racial groups. The audit neglected to consider that where the system failed it did so by placing black defendants in a riskier category, and white defendants in a less risky category. In other words, there was a mechanism in place which was meant to catch problems like this, but it failed because the bias manifested in a way which was different from what the auditing program was looking for.

A similar example, from Arkansas, concerns an algorithmic decision-making system for allocating medical resources which, through an error in design, failed to accurately assess the needs of patients with cerebral palsy or diabetes.[6] Once again, the algorithm's authors insisted that it was working as intended, even as the beneficiaries, and their advocates, could see that something was not right. The problem was only uncovered as a result of litigation brought by Legal Aid of Arkansas on behalf of people who had complained about the cuts to their support.

These examples are illustrative of the range of challenges which can flow from introducing algorithmic decision-making systems to replace or supplement human decision-makers. Some of these problems have been identified, but there are others which likely have yet to be discovered, and may only begin to manifest as these systems are implemented in new contexts. In both cases though, there is an emerging regulatory gap, between systems for accountability, transparency, and engagement which were designed with human decision-makers in mind, and which must now be adapted to deal with the increasing prevalence of automated systems in the decision-making process.

3. The Government's Response

In an effort to address this regulatory gap, on 1 April 2019 the Canadian government's *Directive on Automated Decision-Making* (the *Directive)* took effect.[7] As a key starting point, the *Directive*, for the first time, defines the overall objective of introducing automated decision-making systems, namely in having "more efficient, accurate, consistent, and interpretable decisions made pursuant to Canadian law."[8] It also includes a number of mechanisms for promoting this objective, which generally focus on transparency, auditing, and quality assurance. On the transparency side, the *Directive* includes a requirement that institutions which utilize automated decision-making systems provide clear, prominent and plain-language notices to the public of this fact on their website.[9] The *Directive* also introduces auditing and testing requirements, including for data biases.[10]

Beyond these baseline requirements, the *Directive* introduces a requirement to carry out an "Algorithmic Impact Assessment" prior to the production of an automated decision-making

---

[5] Data & Society, *Algorithmic Accountability: A Primer* (2018) at 5.
[6] AI Now Institute, *Litigating Algorithms* (2018) at 9. A broader description of the lawsuit is also available at: Colin Lecher, "What Happens When An Algorithm Cuts Your Health Care", *The Verge* (21 March 2018), online: <https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy>.
[7] *Directive on Automated Decision-Making*, *supra* note 1.
[8] *Directive on Automated Decision-Making*, *ibid* at s 4.1.
[9] *Directive on Automated Decision-Making*, *ibid* at s 6.2.
[10] *Directive on Automated Decision-Making*, *ibid* at s 6.3.

system, and to publish the results online.[11] The results of this assessment, in terms of the magnitude of the decision's impact and the importance of the rights or interests engaged, leads to a sliding scale of obligations, from baseline requirements for data bias testing and the provision of generalized explanations for common decision results to, at the higher end, requirements for human intervention in the decision-making process, for publication and peer-review, for the provision of a "a meaningful explanation" for negative outcomes, and for Treasury Board approval for the system to operate.[12]

The passage of the *Directive* is a welcome development insofar as it demonstrates that the government is beginning to think critically about the impacts of automated decision-making processes. In other words, we are no longer sleep-walking into the implementation of these new tools. However, while the *Directive* includes some positive aspects, it also leaves a number of major unanswered questions, and fails to provide for a comprehensive regulatory response to the challenges laid out above.

4. Shortcomings of the *Directive*

a. Limited Scope

A significant limitation of the *Directive* are the restrictions on its scope. Robust transparency is the necessary first step to establishing accountability over the use of automated decision-making systems. While the *Directive* certainly provides an important step forward from the current dynamic, under which there is virtually no proactive public disclosure of where and how these systems operate, it also fails to provide the full accounting which the public needs and deserves. A number of offices are wholly excluded from the ambit of the *Directive*, including the Information Commissioner, the Privacy Commissioner, and the Commissioner of Official Languages.[13] The Directive also does not apply to several federal agencies, including the Canada Revenue Agency.

While there are structural reasons for the exclusion of agents of Parliament, much more problematic is the fact that the policy does not apply to "National Security Systems",[14] defined in the *Policy on Management of Information Technology* as any system whose "compromise could undermine the national security of Canada or its partners".[15] Depending on how it is interpreted, this is potentially an extremely broad category, which could apply to any number of highly impactful areas, from immigration and refugee assessments, to airport screening selection, to policing strategies, to mass surveillance, which could be totally excluded from any transparency or accountability mechanisms contained in the *Directive*.

It is also worth noting that the *Directive* does not apply to systems operating in a "test environment". This is defined in a relatively circular manner in Appendix A, leaving it unclear as to whether these tests refer only to internal development and testing, or potentially to more advanced piloting that might include interfacing with and impacting on actual decision-making

---

[11] *Directive on Automated Decision-Making*, *ibid* at s 6.1.
[12] *Directive on Automated Decision-Making*, *ibid* at Appendix C.
[13] *Directive on Automated Decision-Making*, *ibid* at 9.1.
[14] *Directive on Automated Decision-Making*, *ibid* at s 5.4.
[15] Treasury Board of Canada Secretariat, *Policy on the Management of Information Technology* (2007), online: <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=12755&section=html>, at Appendix.

processes.[16] The *Directive* also allows for further exceptions to its requirements to be granted by the Chief Information Officer of Canada, in consultation with the Enterprise Architecture Review Board.[17]

b. Lack of Opportunities for Public Engagement

Perhaps the most serious deficiency of the *Directive* lies in the fact that it does not contain a mechanism for meaningful public engagement, or for those impacted by the implementation of automated decision-making systems, or frontline experts working in the field, to provide feedback on its effects. This outreach is an essential component for building trust in algorithmic systems. In turn, public trust in governance structures is a cornerstone of democracy.

However, beyond the political considerations, there are solid practical reasons for wanting to ensure that communities on the front lines of implementation have a formal mechanism for providing feedback, since they will often be best placed to spot problems in implementation as they occur. Absent a proper mechanism for generating feedback, this expertise will not always filter upward, which may force those communities to pursue costly litigation in order to ensure their concerns are adequately heard. It goes without saying that this option will not be available to everyone, in particular marginalized communities, who are often both the earliest and the most severely impacted by the implementation of automated decision-making systems. The Arkansas and COMPAS cases exemplify both the severe impact of algorithmic decision-making on vulnerable groups, and the importance of providing these communities with a formal avenue for expressing their concerns. In both cases, the individuals impacted knew that something was wrong, even as officials involved in developing and auditing the systems insisted there was no problem.

Absent this vital component, the *Directive* will have limited efficacy in addressing challenges connected to the implementation of automated decision-making systems. Although, as noted above, the introduction of some level of public disclosure is an important step forward, without a formal avenue for expressing their concerns there is a hard limit as to what the public can do with this information. It is also worth questioning how meaningful these notifications are in an age where people are used to disregarding such pro-forma disclosures.[18]

Auditing requirements are also a limited solution, as was well demonstrated by the Arkansas and COMPAS cases. In addition to the inherent challenges with testing these systems, the efficacy of auditing processes is limited by a lack of widely accepted industry standards in this area.[19] Moreover, forms of internal evaluation which focus on each individual process may also face a challenge of "stovepiping", where assessments that look only at the impact or consequences of each individual algorithmic system may miss their broader or combined effect on a particular community.

Even the notion of forcing a human into the loop may be insufficient to alleviate problems arising from algorithmic bias. As early as 1992, it was recognized that algorithmic assessments,

---

[16] *Directive on Automated Decision-Making*, *supra* note 1 at s 5.3 & Appendix A.
[17] *Directive on Automated Decision-Making*, *ibid* at s 8.2.
[18] Margot Kaminski, "The Right to Explanation, Explained" (2019) 34 Berkeley Technology Law Journal at 19-20. See also Lilian Edwards & Michael Veale, "Slave to the Algorithm? Why a 'Right to and Explanation' is Probably not the Remedy You Are Looking For" (2017-2018) 16 Duke Law and Technology Review 18 at 23.
[19] *Algorithmic Accountability*, *supra* note 5 at 8.

even when used purely in a supporting role for human decision-making, had a tendency to undermine that human element, as a result of the "apparently objective and incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities."[20]

5. The Solution

In contrast to the problems noted here, the *Directive* has one very important strength, insofar as there is an automatic review process which must take place every six months.[21] Regular review is a good practice in virtually every aspect of law and governance, but it is particularly important when dealing with novel and emerging areas, where standards are still being developed. While there is a possibility that the government may view these reviews as a mere pro forma box to check, it should use its next review to undertake a robust reconsideration of its approach, and introduce a number of substantive improvements to the *Directive*.

First, and most obviously, the *Directive*'s blanket exclusions should be replaced by specific, harm-based limitations on disclosure. For example, rather than excluding any disclosure at all related to national security systems, the *Directive* could allow for classification of certain aspects of these systems where their release would compromise their efficacy, or otherwise harm key national security interests. This approach is in line with well-established better practice for all right to information and access to information legislation.[22] Even if there were instances where a program were so sensitive that its very existence could not be disclosed, it is difficult to see why that should immunize the system from the regular auditing or human-intervention requirements that are applied elsewhere.

Second, the government should create robust consultation processes connected with their use of automated decision-making systems. This should include, at a minimum, a specific and meaningful procedure to facilitate complaints or feedback from individuals or communities impacted by these systems, and a formal process to raise concerns about biased or otherwise problematic results being returned. More broadly, it may be worthwhile to consider a national consultation on the implementation of automated decision-making systems, to address public concerns and consider public priorities in guiding these systems' implementation. Canada's participation in the Open Government Partnership could provide an interesting model for engagement here, particularly in terms of the Multi-Stakeholder Forum which Canada convened to support dialogue between government, academics, and Canadian civil society.[23] While individual accountability and engagement mechanisms cut against the scalability and efficiency of

---

[20] *Amended Proposal for a Council Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*, at 26, COM(92) 422 final—SYN 297 (Oct. 15, 1992), cited in Edwards & Veale, *supra* note 17 at 27.

[21] *Directive on Automated Decision-Making*, *supra* note 1 at s 1.3.

[22] See, e.g., "Briefing Note Series on Freedom of Expression: The Right to Information" (January 2015), online: *Centre for Law and Democracy* <https://www.law-democracy.org/live/wp-content/uploads/2015/02/foe-briefingnotes-3.pdf>.

[23] "Multi-Stakeholder Forum on Open Government" (13 April 2018), online: *Open Government portal* <https://open.canada.ca/en/multi-stakeholder-forum-open-government>.

automated systems,[24] which is meant to be one of their main benefits, this level of civic engagement in nonetheless necessary to ensure the integrity of these systems, and to facilitate public trust in their use.

Finally, the government should consider creating an independent body to oversee and review its use of automated systems or, barring that, should consider formally delegating this task to an existing independent oversight body, such as the Privacy Commissioner, while boosting that office's funding in order to accommodate the additional workload. This is important in order to ensure that the broader impacts of automated decision-making are considered, in addition to the individual evaluations applied to each system as it is developed.

6. Conclusion

Despite the problems noted in this assessment, the passage of the *Directive* marked an important step forward in Canada's national conversation around the government's use of automated decision-making systems. The basic transparency provisions are a game-changer in supporting important public policy debates in this space, placing Canada ahead of many of its peers. Nonetheless, it is important for Canada not to rest on its laurels, as there is an opportunity to assume a mantel of global leadership in these debates.

Fundamentally, the purpose of automated decision-making systems should never be about replacing human decision-making. Rather, it should be about enhancing human decision-making processes, by boosting their efficiency, accuracy, and consistency. This may not always require a human intervening in every decision. But it is important that the human element of these processes should not be lost. While the auditing and disclosure systems in the *Directive* are an important step toward accountability, the responsible implementation of automated decision-making systems requires a long conversation about what Canadians value in government, and ample opportunity for those most impacted by these systems to express themselves about the challenges and deficiencies they see.

[24] Kroll et al, "Accountable Algorithms" (2017) 165 U Pa L Rev 633 at 639.

THE INCONSENTABILITY OF FACIAL SURVEILLANCE

Evan Selinger and Woodrow Hartzog***

ABSTRACT

Governments and companies often use consent to justify the use of facial recognition technologies for surveillance. Many proposals for regulating facial recognition technology incorporate consent rules as a way to protect those faces that are being tagged and tracked. But consent is a broken regulatory mechanism for facial surveillance. The individual risks of facial surveillance are impossibly opaque, and our collective autonomy and obscurity interests aren't captured or served by individual decisions.

In this article, we argue that facial recognition technologies have a massive and likely fatal consent problem. We reconstruct some of Nancy Kim's fundamental claims in Consentability: Consent and Its Limits*, emphasizing how her consentability framework grants foundational priority to individual and social autonomy, integrates empirical insights into cognitive limitations that significantly impact the quality of human decision-making when granting consent, and identifies social, psychological, and legal impediments that allow the pace and negative consequences of innovation to outstrip the protections of legal regulation.*

We also expand upon Kim's analysis by arguing that valid consent cannot be given for face surveillance. Even if valid individual consent to face surveillance was possible, permission for such surveillance is in irresolvable conflict with our collective autonomy and obscurity interests. Additionally, there is good reason to be skeptical of consent as the justification for any use of facial recognition technology, including facial characterization, verification, and identification.

---

* Evan Selinger is a Professor of Philosophy at Rochester Institute of Technology.

** Woodrow Hartzog is Professor of Law and Computer Science at Northeastern University School of Law and Khoury College of Computer Sciences. The authors would like to thank Kyle Berner for his excellent research assistance.

I. INTRODUCTION

"Surveillance" is an ominous word. In the post-Snowden world, it evokes Orwellian watchers who observe our every move, as persistent as they are powerful. Given the strong reactions the term can evoke, why hasn't greater resistance manifested against surveillance threats? An important reason is that surveillance technology is deployed in ways that make us feel comfortable with, not creeped out by, the algorithms and people observing us.[1] Facebook, for example, is designed to be an environment that feels so intimate that users focus on sharing information with friends without thinking about "surveillance capitalism" and all of the data the company collects, analyzes, and monetizes on the back end.[2] At airports and concerts, the experience of using facial recognition technology, a tool that is used for racial profiling and tracking in China and to scan the streets of Russia for "people of interest," can feel like a godsend, saving us and everyone else who socially conforms from waiting in long frustrating lines.[3] The more familiar and beneficial a surveillance technology like facial recognition seems, the easier it is for technology companies, government agencies, and entrepreneurs to create conditions for widespread passive acceptance.

Normalization, which involves treating facial recognition technology as a mundane part of the machinery that is necessary for powering a complex digital society, and function creep, which entails incrementally expanding how the technology is used, mask harms to individual and collective autonomy. They make it easy for surveillers to operate within a permissive regulatory regime: one that has porous boundaries between the government and the private sector, and treats consent as the basis for authorizing permission for watching, tagging, tracking, and sorting.[4] Even when our consent is obtained through questionable means, perhaps nudged by dark patterns and hidden options, many of us

---

1. *See* Evan Selinger, *Why Do We Love To Call New Technologies "Creepy"?*, SLATE (Aug. 22, 2012), https://slate.com/technology/2012/08/facial-recognition-software-targeted-advertising-we-love-to-call-new-technologies-creepy.html.

2. Evan Selinger, *Facebook Fabricates Trust Through Fake Intimacy*, MEDIUM (Jun. 4, 2018), https://medium.com/s/trustissues/facebook-fabricates-trust-through-fake-intimacy-b381e60d32f9.

3. Ian Sample, *What is facial recognition-and how sinister is it?*, THE GUARDIAN (July 29, 2019), https://www.theguardian.com/technology/2019/jul/29/what-is-facial-recognition-and-how-sinister-is-it.

4. For more on normalization and function creep, *see* BRETT FRISCHMANN AND EVAN SELINGER, RE-ENGINEERING HUMANITY (2018).

will say yes when companies ask for it while engaging in surveillance or surveillance-related activities.[5] With limited alternatives to choose from and barriers to collective action that impede creating new, less surveillance intensive options, assenting to surveillance seems like the most rational "choice" for avoiding the penalties that come from being an opt-out outlier while accruing whatever take-it-or-leave-it benefits are offered by the consent-seeker, however meager they may be.[6]

The law has long struggled with problems associated with consent. In *Consentability: Consent and Its Limits*, Nancy Kim provides a promising path forward by integrating legal and ethical scholarship on consent with scientific inquiry into humanity's predictable irrationality. Drawing from these interdisciplinary resources, she constructs a new consentabilty framework and applies it to difficult cases: assisted suicide, body modification (from cosmetic surgery to RFID chip implants), bodily integrity exchanges (sexual services, surrogacy, and organ sales), and experimental activities (such as traveling to Mars and becoming cryopreserved).

In this article, we draw upon Kim's work along with our previous research on surveillance and privacy theory to make one simple point: facial recognition technologies probably have a fatal consent problem. After reviewing some of Kim's main ideas, we will apply aspects of her framework to explore how facial recognition technologies generally, and face surveillance specifically, affects us in ways that are difficult for most people to appreciate.

When we use the term face surveillance, we mean the use of facial recognition technologies and faceprint or name-faceprint databases to monitor behavior, identify people, or gain insight or information for the purposes of influencing, managing, directing, or deterring people. Examples include real-time observation, tracking, and identifying people in airports, retail stores, and public parks, as well as using faceprints and algorithms to identify and analyze people in stored photos and videos for law enforcement, commercial, and marketing purposes. The Future of Privacy Forum conceptualized instances of "identification: one to many" as situations where software tries to determine who an

5. For more on the conflicts between design and valid consent, *see* WOODROW HARTZOG PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES 5 (2018).

6. *See* Frischmann and Selinger, *supra* note 4.

unknown person is, and "unique persistent identifiers," which are cases where algorithms try to determine what someone is doing "in a limited context, not linked to other personal identifiable information?"[7] We also use the terms "facial detection," which are instances of software trying to determine if a face can be found in a picture, and "facial characterization," which are situations where algorithms code assumptions about faces, such as emotions people might be experiencing.

We argue that valid consent is not possible for face surveillance in many of its current and proposed applications because of its inevitable corrosion of our collective autonomy, to say nothing of the dubious validity of individual consent in these contexts.[8] Additionally, we argue that some forms of characterization are inconsentable due to collective autonomy problems and are at least vulnerable to defective consent. Even "1:1 facial identification" features are highly subject to defective consent and should be highly scrutinized. Only facial detection tools ("is this a face?") seem entitled to the benefit of the doubt because they are not used to persistently track, identify, or manipulate people.

One reason consent to facial recognition is highly suspect is that people do not and largely cannot possess an appropriate level of knowledge about the substantial threats that facial recognition technology poses to their own autonomy.[9] Additionally, the framing of this debate around the amorphous concept of individual "privacy" has hidden unjustifiable risks to two of the most important values implicated by facial recognition: obscurity and collective autonomy. Even if some people withhold consent for face surveillance, others will inevitably give it. Rules that facilitate this kind of permission will normalize behavior, entrench organizational practices, and fuel investment in technologies that

---

7. Brenda Leong, *FPF Releases Understanding Facial Detection, Characterization, and Recognition Technologies and Privacy Principles for Facial Recognition Technology in Commercial Applications*, FUTURE OF PRIVACY FORUM (Sept. 20, 2018), https://fpf.org/wp-content/uploads/2019/03/Final-Privacy-Principles-Edits-1.pdf.

8. In addition to drawing from our own research and prior collaborations, our approach to analyzing consent will integrate insights from Neil Richards and Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. U. L. REV. 1461 (2019).

9. The entire field of behavioral economics is built around the idea that people have limited knowledge and capacity as decisionmakers. *See, e.g.,* DANIEL KAHNEMAN, THINKING FAST AND SLOW 4 (Farrar, Straus and Giroux ed., 2011); DAN ARIELY, PREDICTABLY IRRATIONAL (HarperCollins ed., 2008); CASS SUNSTEIN AND RICHARD THALER, NUDGE (2008).

will result in a net increase of surveillance. Expanding a surveillance infrastructure will increase the number of searches that occur which, in itself, will have a chilling effect over time as law enforcement and industry slowly but surely erode our collective and individual obscurity.

Building an infrastructure to facilitate surveillance will also provide more vectors for abuse and careless errors. No one is perfect, and the more requests for permission to surveil that are made the more harm from mistakes and malice will exist. Additionally, the larger and more entrenched facial recognition infrastructure becomes, the more opportunities exist for law enforcement to bypass procedural rules on searches to obtain information directly from industry. For example, if the government were prohibited from directly using facial recognition technologies, it could purchase people's location data obtained from facial recognition technology (and thus linked to their identities) from private industry. Procedural rules wouldn't address the true harm of these technologies without further prohibitions to prevent end-runs around the aims of a restriction.

We conclude this article with the argument that to defend against these dangers, lawmakers should pursue strong policy measures beyond procedural protections such as warrant requirements and informed consent frameworks. At a minimum, lawmakers should immediately enact moratoriums to prevent entrenchment of and dependence on facial recognition systems before they can be properly considered by lawmakers and society. In all areas where consentability conditions cannot be met, and procedural rules and compliance frameworks for government and industry will facilitate an outsized harm and abuse relative to their gains, facial recognition technology should be outright banned.

II. CONSENTABILITY AND INVALID CONSENT

Consent is a foundational concept in the American law. As one of us wrote with Neil Richards,

> We live in a society that lionizes individual choice in the many social roles we play every day, whether as consumers, citizens, family members, voters, lovers, or employees. Consent reinforces fundamental cultural notions of autonomy and choice. It transforms the moral landscape between people and makes the otherwise impossible possible.[1] It is essential to the exercise (and waiver) of fundamental constitutional rights, and it is at the essence of political freedom, whether we are

> talking broadly about a "social contract" or making political choices for individual candidates and referenda in the voting booth.[10]

Morally and legally, consent involves the "'intentional transfer of rights and obligations between parties,' which transforms the moral landscape between them and makes the otherwise impossible possible."[11]

Kim noted that "[c]onsent in the law is typically viewed as a conclusion, an all-or-nothing concept where the actions of the parties are considered objectively and statically."[12] The problem with this, Kim argued, is that "[t]his conception provides no guidance regarding which acts should be consentable."[13] According to Kim, "while the requirement of consent recognizes the value of autonomous decision-making, the *validity* of consent hinges upon the context in which it is given and the dynamic unleashed by both parties."[14] This means that valid consent is not only suspect in some contexts, but not even possible. She labels this concept regarding the circumstances under which consent can be valid "consentability."

In Kim's framework, consentability revolves around two requirements. First, an individual must be able to validly consent to a proposed activity. This means that they can intentionally manifest consent, possess the requisite knowledge in light of the motive for consenting, and exercise their volition to do so. Second, the social benefits of the activity must outweigh the social harms. In both cases, Kim maintains there is a range of fundamental yet hierarchically differentiable interests that the liberal state should safeguard: equality, justice and due process, public safety, democracy, free market capitalism, the right to bodily integrity, freedom of movement, civil and political rights, and property rights. At their core, Kim contends all these interests are expressions of autonomy, which she argues is a primary societal value. Since people can be born into a range of life-impacting circumstances that are beyond their control, the fairest way to foster and protect everyone's autonomy is to configure a social order that promotes liberty for all citizens. While individuals have

---

10. Neil Richards and Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. U. L. REV. 1461, 1462 (2019).
11. *Id.* at 1462, 1468.
12. NANCY KIM, CONSENTABILITY: CONSENT AND ITS LIMITS 3 (2019).
13. *Id.*
14. *Id.*

autonomy interests at the personal level, Kim also identifies collective autonomy interests, which she defines as "the interest that all members of a society have in a particular right." From this structural perspective, if a clash occurs over comparable autonomy interests, Kim insists that "the collective autonomy interest prevails over the individual autonomy interest."[15]

At the individual level, Kim identified three essential features underlying legal determinations of consent. They are "an intentional manifestation of consent, knowledge, and volition/voluntariness."[16] Ideally, a person should not agree to an offer unless she understands what it entails, freely chooses to enter into the agreement, and demonstrates her agreement through clear words or deeds. In the real world, however, each condition is challenging. Voluntariness is vexing because real people, unlike hypothetically postulated rational actors, are bound by so many constraints that "no human being is truly or ideally autonomous all the time."[17] Clear affirmation is debated because the standard is context dependent. For example, Kim endorses some transactions requiring the consenting party to sign once at the end of a contract. However, she objects to the one-and-done practice being used in other circumstances, such as manifesting "consent to a bodily integrity contract where the consenter agrees to transfer his kidney."[18] While these are daunting complications, Kim deems the knowledge condition to be the hardest one to satisfy. This is because people can make poor decisions not only when they lack pertinent information, but also when they have access to all of the relevant details.

The problem of missing information is self-evident. But why doesn't having enough of it suffice for making informed decisions? It is because the quality of information matters. In order for information to be useful, it must be "understandable and salient."[19] Unfortunately, U.S. contract law exacerbates the problem. It incentivizes creating contracts that use jargon and provide overwhelming amounts of detail.[20] As a result, online user agreements regularly minimize the consent seeker's liability by hiding risks in plain sight.

---

15. NANCY KIM, CONSENTABILITY: CONSENT AND ITS LIMITS 84, 88 (2019).
16. *Id.* at 9.
17. *Id.* at 55.
18. *Id.* at 122.
19. NANCY KIM, CONSENTABILITY: CONSENT AND ITS LIMITS 125 (2019).
20. *See, e.g.,* Frishmann & Selinger, *supra* note 4; Neil Richards and Woodrow Hartzog, The Pathologies of Digital Consent, 96 WASH. U. L. REV. 1461, 1484 (2019).

To illustrate this problem, Kim declares that "a company that creates a product that records a person's conversations and collects their images should not be able to justify those actions by claiming that its customers consented by clicking 'agree' to the company's terms and conditions."[21]

To determine how to communicate a risky opportunity without rendering consent illegitimate, Kim turns to cognitive science and behavioral economic research on bounded rationality and the dual-process model of human cognition. In accordance with leading dual-process theorists, Kim maintains that human decision-making capacity is flawed in many ways, often in ways that we are unaware of. For example, we may not know whether our decisions are guided by the deliberative or intuitive cognitive system, if our decisions are impaired by heuristic techniques laden with cognitive biases, if we are self-sabotaging by misperceiving irrational decisions as rational ones, and if we are being swayed by misleading or manipulative information. From this perspective, people may make choices they later regret due to flawed heuristics like representative, anchoring, and availability; cognitive biases like overconfidence, optimism, and confirmation; heated emotional and physical states; or an inclination towards social conformity.[22]

While being attuned to cognitive limitations is necessary for formulating communication criteria that satisfies the knowledge condition, it is also insufficient. When consent is sought, the quality of information provided must be calibrated to adjust for two things: how much risk the transaction poses to individual and collective autonomy, and how trustworthy the consent-seeking parties are. Kim thus tailors her consentability framework on a sliding scale of consent standards. The greater the risk to autonomy, the more she believes a person is entitled to understand. For extremely risky situations, such as ones that could lead to "permanent disfigurement," Kim argues the "conditions of consent must be established with absolute certainty, the equivalent of the judicial standard 'beyond a reasonable doubt.'"[23]

By linking risk-level to the quality of consent-seeking disclosures, Kim derives a basis for demarcating valid from invalid consent at the individual level. She argues that consent is invalid if "the threat to autonomy interest outweighs the robustness of the

21. KIM, *supra* note 12, at 119.
22. *Id.* at 13.
23. Nancy Kim, Consentability: Consent and Its Limits 83 (2019).

consent conditions."[24] This means that if a transaction poses a great threat to autonomy and the consent conditions are not commensurate with the risk, valid consent cannot be given.

Although it might seem that consent must be either valid or invalid since an offer either can meet or fall short of the consentability standard, things are actually more complicated. An offer accepted under deficient consentability conditions results in one of two outcomes. Either the transaction transpires without genuine consent being given or else the offer is accepted through "defective consent." Kim characterizes this outcome as the "purgatory between valid consent and non-consent."[25] Kim's paradigm case of defective consent is a patient in an emergency situation agreeing to a medical procedure out of fear that failing to do so will pose high-level risks to her autonomy. In this instance, the patient is not acting in a truly voluntary manner. Even when professional standards nevertheless allow her to proceed with the procedure, Kim maintains that contractual bargaining should not transpire that includes terms that limit "the liability of the surgeon for malpractice nor require the patient to agree to mandatory arbitration in the event of a dispute."[26]

III. FACIAL RECOGNITION TECHNOLOGY DYSTOPIA

Consentability contains a passage about technology-induced change that is so bleak, it is worth quoting at length.

> Technology will continue to push the boundaries of what society thinks is acceptable. In some cases, the changes will be gradual, occurring first on the fringes of society and undetected by the public. . . . Sometimes the changes will go undetected because they are not visible or obvious to most people. As Lori Andrews observed in the context of genetics policy, "When technologies are introduced incrementally and policies are adopted in small units to deal with a few isolated issues, there is less opportunity to stimulate a social debate about whether we are moving in a direction in which we want to go." Companies, skilled in the art of marketing and sales, may try to manipulate the public and intimidate lawmakers into accepting products and services which degrade, rather than enhance, social relations. Legislatures will be indifferent or reluctant to act until there is some sort of social outcry or

24. *Id.* at 81.
25. *Id.* at 132.
26. *Id.*

> the impact on society is too great to ignore. The law will arrive too late, after social norms have already been established and when it is much more difficult to reverse society's course.[27]

Before showing how Kim's consentability framework can be applied to the facial recognition technology debates, we will sketch the outline of dystopian future. The scenario is a thought experiment about a possible world where the dire risks posed by facial recognition technology poses are realized. The transition from the present world to this hypothetical future could occur due to structural problems like the ones Kim outlines in the above passage.

Much of the discussion about the immediate and short to medium term problems with facial recognition technology focuses on the harm that could occur if the technology continues to produce inaccurate results.[28] Law-abiding people could be put on government watchlists, deprived of due process in court, prevented from accessing places they should be allowed to enter, and questioned or detained by law enforcement. Government and industry could deny people access to their assets, deprive them of job opportunities, and mischaracterize their identities and behaviors. While everyone is vulnerable to these harms, false positives and negatives disproportionately affect minorities, especially people of color.[29] These discussions also emphasize that the law poses few restrictions on facial recognition technology. Furthermore, there is little transparency about how facial recognition technology is used as we can see from the fact that state legislatures are not required to openly debate and approve (i.e., consent) using driver's license photos for government facial recognition databases.[30] Finally, internal policies for the

---

27. NANCY KIM, CONSENTABILITY, CONSENT AND ITS LIMITS S 118-119 (2019).

28. *See, e.g.,* Sahil Chinoy, *The Racist History Behind Facial Recognition*, N.Y. TIMES (July 10, 2019), https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html; Steve Lohr, *Facial Recognition is Accurate, if You're a White Guy*, N.Y. TIMES (Feb. 9, 2018), https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html; Natasha Singer, *Amazon's Facial Recognition Wrongly Identifies 28 Lawmakers, A.C.L.U. Says*, N.Y. TIMES (July 26, 2018), https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html.

29. *See* Joy Boulamwini, *When the Robot Doesn't See Dark Skin*, N.Y. TIMES (June 21, 2018), https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html.

30. Drew Harwell, *FBI, ICE find state driver's license photos are a gold mine for facial-recognition searches*, WASH. POST (July 7, 2019),

government using facial recognition technology are not standardized.

Over time, advances in facial recognition technology might eliminate all kinds of errors. Unfortunately, more accurate versions of the technology pose even greater dangers because the problems with facial surveillance are fundamental and unique. Evan Greer contends, "Biometric surveillance powered by artificial intelligence is categorically different than any surveillance we have seen before. It enables real-time location tracking and behavior policing of an entire population at a previously impossible scale."[31] The technology can be used to create chill that routinely prevents citizens from engaging in First Amendment protected activities, such as free association and free expression. They could also gradually erode due process ideals by facilitating a shift to a world where citizens are not presumed innocent but are codified as risk profiles with varying potentials to commit a crime. In such a world, the government and companies alike will find it easy to excessively police minor infractions, similar to how law enforcement already uses minor infractions as pretexts to cover up more invasive motives.[32] Surveillance tools bestow power on the watcher. Abuse of the power that was once localized and costly could become systematized, super-charged, and turnkey. Companies could expand their reach of relentless and manipulative marketing by peddling their wares over smart signs that display personalized advertisements in public spaces. And as more emotional states, private thoughts, and behavioral predictions are coded from facial data, people will lose more and more control over their identities. They could be characterized as belonging to groups that they don't identify with or don't want everyone knowing they belong to. And while schools might monitor students more intensely and make the educational environment more like a prison, bad actors will have opportunities to create even more general security problems through hacking and scraping.

---

https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/.

31. Evan Greer, *Don't Regulate Facial Recognition. Ban it.*, BUZZFEED NEWS (July 18, 2019), https://www.buzzfeednews.com/article/evangreer/dont-regulate-facial-recognition-ban-it.

32. *See* Angela Caputo, *Berwyn Police Rack up Citations with Questionable DUI Checkpoints*, CHI. TRIB. (Sept. 20, 2015), https://www.chicagotribune.com/investigations/ct-berwyn-dui-checkpoints-met-20150920-story.html.

How might this social transformation occur? With the law lagging behind innovation and an existing legacy of name-face databases ripe for plug-and-play expansion, the perceived advantages of easily and cheaply analyzing biometric faceprints that link our on- and off-line lives could drive widespread adoption. As this happens, people could get used to thinking of facial recognition technology as the go-to solution for solving all kinds of problems throughout society. Tired of remembering and entering in a passcode to unlock your phone? Try facial recognition. Long lines boarding a plane? Maybe facial recognition could help. Not sure who's knocking at your door? Facial recognition could tell you. Missing your child while they're at summer camp and want to watch them play? Facial recognition to the rescue! And so on.

Patching social problems with technological solutions is easier than mustering the will to solve harder issues around inequality, education, and opportunity. The drumbeat of security stokes fear. And enhancing convenience is a powerful motivating force in American life. Consequently, it won't be reasonable to expect most people to grasp that they should summon the political will to push back against incremental buildup of negative effects that initially concentrate the worst outcomes on people of color and activists. Immediate gratification, abstract perceptions of risk, and certain harm is a recipe for doom.

IV. THE FRAMING PROBLEM: OBSCURITY, NOT PRIVACY OR ANONYMITY

To apply Kim's insights to the debate over facial recognition technology, it is useful to begin by leveraging a concept from the literature on cognition that she relies upon: *framing effects*. Word choice can have a framing effect because how options and issues are presented can impact how people perceive risks and what solutions they propose. For example, since research into the cognitive bias of loss aversion suggests that people tend to perceive losses as more significant than gains, it matters whether doctors describe a surgical procedure as having a 90% success rate or a 10% failure rate.[33]

33. Erving Goffman, Frame Analysis 7 (Harper Colophon ed., 1974); Robert D. Benford & David A. Snow, *Framing Processes and Social Movements: An Overview and Assessment*, 26 ANN. REV. SOC. 611, 614 (2000); Dennis Chong & James N. Druckman, *Framing Theory*, 10 ANN. REV. POL. SCI. 103, 104 (2007); Laura E. Drake & William A. Donohue, *Communicative Framing Theory in Conflict Resolution*, 23 COMM. RES. 297, 300 (1996); Daniel Kahneman & Amos Tversky, *Choices, Values, and Frames*, 39 AM. PSYCHOLOGIST 341, 341 (1984); Deborah Tannen, *What's in a Frame? Surface*

The debates over facial recognition technology, like other debates over surveillance, are marred by the fact that they are framed around the concepts of "privacy" and "anonymity" instead of "obscurity."[34] The harm from surveillance is often described as loss of privacy.[35] But the concept of privacy is famously amorphous. It can mean almost anything from secrecy to intimacy to control to "the right to be let alone."[36] With respect to surveillance, people often make the argument that as long as you're in "public," people can already see you; since it is not reasonable to ask people to avert their eyes in public, you allegedly have no privacy in accessible spaces.[37] Others make the argument that they don't fear surveillance as a privacy threat because they have "nothing to hide."[38] These arguments either reduce privacy to secrecy and assume that only things that are completely stowed away are worthy of protection, or else myopically frame privacy as a concern for individuals, not society writ large.

At least initially, framing surveillance harms in autonomy terms is also problematic. This is because the concept of autonomy can be stretched in an almost limitless fashion. Jeb Rubenfeld writes:

> What, then, is the right to privacy? What does it protect? A number of commentators seem to think that they have it when they add the word 'autonomy' to the privacy vocabulary. But to call an individual 'autonomous' is simply another way of saying that he is morally free, and to say that the right to privacy protects freedom adds little to our understanding of the doctrine. To be sure, the privacy doctrine involves the 'right to make choices and decisions,' which, it is said, forms the 'kernel' of autonomy. The question, however, is which

---

Evidence for Underlying Expectations, in Framing in Discourse 137 (Deborah Tannen ed., 1979); Amos Tversky & Daniel Kahneman, *The Framing of Decisions and the Psychology of Choice*, 211 SCIENCE 453, 453 (1981).

34. *See, e.g.,* Joseph Kupfer, *Privacy, Autonomy, and Self-Concept*, 24 AM. PHIL. Q. 81,81 (1987); Louis Henkin, *Privacy and Autonomy*, 74 COLUM. L. REV. 1410, 1419 (1974).

35. *See* Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L. J. 1131, 1131 (2011).

36. *See* DANIEL SOLOVE, UNDERSTANDING PRIVACY 13 (First Harvard Univ. Press eds., 2008).

37. For an exploration and rebuttal of this argument, *see* Woodrow Hartzog, *The Public Information Fallacy*, 99 B.U. L. REV. 459, 461 (2019).

38. *See* Daniel J. Solove, *"I've Got Nothing to Hide" and Other Misunderstandings of Privacy*, 44 SAN DIEGO L. REV. 745 (2007).

choices and decisions are protected?[39]

While surveillance certainly implicates Kim's twin foci of individual and social autonomy, the concept of autonomy is likely too broad to meaningfully and consistently resonate with people who are making decisions that would put it at risk. In the context of facial recognition technology, autonomy, like privacy, needs a better, more specific, framing. We propose framing surveillance issues generally, and facial recognition specifically, as a loss of "obscurity," a diminution that clearly detracts from many of the goods that autonomy is valued for enabling.

To briefly summarize key points from our extensive prior research, the concept of obscurity concerns transaction costs—the ease or difficulty of finding information and correctly interpreting it.[40] The harder it is to locate information or reliably understand what it means in context, the safer, practically speaking, the information is. Safety is a matter of probability, not certainty, since a range of factors can change transaction costs. Examples of such factors include advances in technological capabilities, the democratization of technological functions, and advances in data science. For much of history, obscurity has been protected by what Harry Surden calls "structural constraints."[41] These are not legal protections, they are technological limitations such as a lack of easy to use, inexpensive, and accurate means of identifying us, tracking our movements, behaviors, and communications, and inferring our thoughts and emotions. Structural constraints may also be biological. For instance, the fact that the human cognitive and perceptual systems can only make sense of and store limited amounts of information without technological aid. While the transaction costs imposed by warrant requirements, encryption software, and other strategies provide some obscurity protections, they are of limited value in a society that rules out privacy protections in public and when information is disclosed to third parties (e.g., the Third Party Doctrine).[42] They are also limited

39. Jed Rubenfeld, *The Right of Privacy*, 102 HARV. L. REV. 737, 750-52 (1989).

40. Evan Selinger & Woodrow Hartzog, *Obscurity and Privacy*, *in* ROUTLEDGE COMPANION TO PHILOSOPHY OF TECHNOLOGY (Joseph Pitt and Ashley Shew, eds. Forthcoming 2014); Woodrow Hartzog & Evan Selinger, *Surveillance as Loss of Obscurity*, 72 WASH & LEE L. REV. 1343 (2015); Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 CALIF. L. REV. 1 (2013); Woodrow Hartzog & Frederic Stutzman, *Obscurity By Design*, 88 WASH. L. REV. 385 (2013).

41. Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605 (2007).

42. Orin S. Kerr, *The Case for the Third-Party Doctrine*, 107 MICH. L. REV. 561 (2009) (providing a defense of the third-party doctrine).

because our society fundamentally does not view privacy in terms of nuanced categories. like select publics or private publics, where information is meant to be disclosed to some audiences but not everyone, rather than blunt ones like anonymity, which presuppose that nobody knows who you are.

In order for people to be capable of giving valid consent to a range of surveillance practices, including facial recognition, they need to have a better understanding of how they rely on obscurity to protect their privacy. By taking obscurity for granted, they miss how it fosters individual autonomy. Obscurity enables people to establish meaningful and intimate relationships because it allows us to selectively disclose information and share different aspects of our identity in different contexts.[43] Obscurity enables us to develop intellectually and emotionally by giving us breathing room to embrace risks and make mistakes without the stigma of being forever associated with failures and fads.[44] Obscurity enables citizens to participate in democracy by allowing them to confidently engage in political activities without worrying about recriminations from the government.

However, such appreciation means little on its own. What good is recognizing the value of obscurity if it is unobtainable? Consequently, this understanding needs to be bolstered by substantial changes to the privacy regulatory regime that provide meaningful obscurity protections. At present, neither a great obscurity awakening, nor a regulatory obscurity revolution are likely; both entail too much of a departure from entrenched theories and practices.

V. FACIAL RECOGNITION TECHNOLOGY: INDIVIDUAL CONSENT AND COLLECTIVE AUTONOMY

Should facial recognition surveillance be consentable? By appealing to Kim's framework to answer this question, we must ask whether it is possible to validly consent to the proposed activity, and whether social harms caused by the activity outweigh its social benefits. It seems unlikely that someone could give valid consent to most forms of facial surveillance because the context in which such consent would be sought frustrates the pre-conditions for meaningful decision-making. In order for consent to data and surveillance practices to be knowing and voluntary, at least three

43. *See* ERVING GOFFMAN, THE PRESENTATION OF SELF IN EVERYDAY LIFE (1956).

44. For an exploration on the importance of privacy for "play" and human flourishing, *see* Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904 (2013).

pre-conditions should exist: (1) such a request should be infrequent, (2) the harms to be weighed must be vivid, and (3) there should be incentives to take each request for consent seriously.[45] If the requests for consent are too frequent people will become overwhelmed and desensitized. This renders them susceptible to user interfaces and dense, confusing, turgid privacy policies that are designed to exploit their exhaustion to extract consent. If the harms are framed in terms of abstract notions of privacy and autonomy or the possibility of abuse is too distant to be readily foreseeable, then people's cost/benefit calculus may be corrupted by an inability to take adequate stock of the risks. Finally, if the risk of harm is distributed over the course of many different decisions—as is common with loss of obscurity through surveillance—people will lack the proper incentive to take each request for consent seriously. After all, no single decision represents a significant threat. Instead, society is exposed to death by a thousand cuts, with no particular cut rising to the threat level where substantive and efficacious dissent occurs.

In the case of facial recognition technology things are further complicated by the fact that the public is routinely given seemingly good reasons to believe that the social benefits caused by consenting to surveillance would outstrip any social harms. As we previously described this illusory worldview:

> From this perspective, you'll never have to meet a stranger, fuss with passwords, or worry about forgetting your wallet. You'll be able to organize your entire video and picture collection in seconds—even instantly find photos of your kids running around at summer camp. More important, missing people will be located, schools will become safe, and the bad guys won't get away with hiding in the shadows or under desks. Total convenience. Absolute justice. Churches completely full on Sundays. At long last, our tech utopia will be realized.[46]

But many of these touted benefits are meager, incremental improvements that could likely be approximated through less dangerous means. For example, facial recognition is being deployed to streamline the hassle associated with paper boarding

---

45. *See* Neil Richards and Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. U. L. REV. 1461, 1466 (2019).

46. Woodrow Hartzog & Evan Selinger, *Facial Recognition is the Perfect Tool for Oppression*, MEDIUM (Aug. 2, 2018), https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66.

passes, cash and debit cards, and passcodes and fingerprint access.[47] But these technologies already worked reasonably (or exceptionally) well. The legitimately compelling benefits, such as finding missing people and keeping people safe, would require large, promiscuous databases working with interconnected and ubiquitous sensors making a mind-bogglingly large number of fraught algorithmic decisions. Such an infrastructure would extract a massive toll on our freedoms, civil liberties, and autonomy. Setting up this infrastructure also intrinsically incentivizes its use due to the sunk cost fallacy, a cognitive bias emphasized by the cognitive science literature that Kim discusses.[48] The sunk cost fallacy is the tendency for humans continue down a particular course once they have made significant investment in it. Spending all the resources required for getting the infrastructure built and stoking expectations that the infrastructure is required for social progress would therefore make it hard to change course and accept the reality that previous resources could have been better spent.

The harms of facial surveillance are legion. The mere existence of facial recognition systems, which are often invisible, harms civil liberties because people will act differently if they suspect they're being surveilled.[49] Even legislation that promises stringent protective procedures won't prevent chill from impeding crucial opportunities for human flourishing by dampening expressive and religious conduct. Warrant requirements for facial recognition will merely set the conditions for surveillance to occur, which will normalize tracking and identification, reorganize and entrench organizational structure and practices, and drive government and industry investment in facial recognition tools and infrastructure.

Facial recognition technology also enables a host of other abuses and corrosive activities, many of which we outlined in the

---

47. *See* Brian Feldman, *Replacing Touch ID With Face ID is a Worse Idea Than You Think*, N.Y. INTELLIGENCER (Sept. 12, 2017), http://nymag.com/intelligencer/2017/09/replacing-touch-id-with-face-id-is-worse-than-you-think.html; Betsy Isaacson, *Facial Recognition Systems Turn Your Face Into Your Credit Card, PIN, Password*, HUFFPOST (July 19, 2013), https://www.huffpost.com/entry/facial-recognition-credit-card_n_3624752; Gregory Wallace, *Instead of the Boarding Pass Bring Your Smile to the Airport*, CNN (Sept. 10, 2018), https://www.cnn.com/travel/article/cbp-facial-recognition/index.html.

48. *See* Jamie Ducharme, *The Sunk Cost Fallacy is Ruining Your Decisions. Here's How*, TIME (July 26, 2018), https://time.com/5347133/sunk-cost-fallacy-decisions/.

49. Neil Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1935 (2013).

previous section.

- Disproportionate impact on people of color and other minority and vulnerable populations.
- Due process harms, which might include shifting the ideal from "presumed innocent" to "people who have not been found guilty of a crime, yet."
- Facilitating harassment and violence.
- Denial of fundamental rights and opportunities, such as protection against "arbitrary government tracking of one's movements, habits, relationships, interests, and thoughts."
- The suffocating restraint of the relentless, perfect enforcement of law.
- The normalized elimination of practical obscurity.
- Digital epidermalization and applied junk science (e.g., digital phrenology).
- The amplification of surveillance capitalism.
- Security vulnerabilities.

Finally, even assuming that an individual could consent, facial recognition systems inevitably will lead to unacceptable harm to our collective autonomy. In a democracy, it is reasonable to expect that many people will put greater weight on the costs and benefits of a particular decision that are relevant to them and people like them. Such is the pull of tribalism and privilege, which bias decision-making much like the compromising factors that Kim emphasizes. In practice, this means if citizens are not members of minority communities, they might not be sufficiently concerned with how their gain from facial recognition comes at other people's expense. Addressing this hidden cost, Chris Gillard aptly states:

> Until we can come to better terms with the disparate impacts of privacy harms, the privileged will continue to pay for luxury surveillance, in the form of Apple Watches, IoT toilets, quantified baby products, Ring Doorbells, and Teslas, while marginalized populations will pay another price: Surveillance, with the help of computer data, deployed against them—in the form of ankle bracelets, license plate readers, drones, facial recognition, and cell-site simulators. As one group pays to be

> watched, other groups continue to pay the price for being watched.[50]

Over time, when majority groups consent to offers that are cost-benefit justified for themselves, large-scale social transformation can result that compromises the autonomy interests of marginalized groups. The end result is likely a society that won't be able to provide an adequate base level of autonomy protections for all citizens. For if marginalized groups come to experience the pervasive chill of having not just their public movements but also their identities (e.g., gay-identifying algorithms) and mental states (e.g., emotion detection) monitored—then the rest of society isn't justified in making choices that lead to this outcome. The end result would be the unraveling of obscurity, and with it, the erosion of democratic legitimacy through tyranny of the majority—an outcome that Kim characterizes as unjust by assigning primacy to collective autonomy in her framework.

VI. CONCLUSION: MORATORIA AND BANS

When Kim considers bans in *Consentability*, she approaches the issue through the framing of paternalism to inquire into the liberties the government is justified in curtailing. For example, she argues that it should not be consentable to smoke tobacco or marijuana in public due to the adverse harm it can cause to third parties, but junk food should only be more restrictively regulated, not banned.[51] Bans, however, are not limited to expressions of state power. In both principle and practice, they also can be restrictions upon it.

To that end, an unexpected shift in governance has begun. U.S. cities have started banning government agents from using facial recognition technology.[52] Statewide moratoriums on government agents are being considered too.[53] Bans, whether temporary or permanent, are extremely rare in U.S. governance because lawmakers and policy advocates often make three core

---

50. Chris Gilliard, *Privacy's Not an Abstraction*, FAST COMPANY (Mar. 25, 2019), https://www.fastcompany.com/90323529/privacy-is-not-an-abstraction.

51. NANCY KIM, CONSENTABILITY: CONSENT AND ITS LIMITS 168-171 (2019).

52. Caroline Haskins, *Oakland Becomes Third U.S. City to Ban Facial Recognition*, VICE MOTHERBOARD (July 17, 2019), https://www.vice.com/en_us/article/zmpaex/oakland-becomes-third-us-city-to-ban-facial-recognition-xz.

53. Steve LeBlanc, *Mass. Lawmakers Aim to Block Facial Recognition Technology*, BOSTON 25 NEWS (June 22, 2019), https://www.boston25news.com/news/mass-lawmakers-aim-to-block-facial-recognition-technology/960520513.

presumptions about regulation. The first is that extreme fears about new technologies should be viewed as over-reactions that parallel previous panics about technologies that society effectively adapted to, such as the automobile, radio, and television.[54] The second is that all dual-use technologies should be integrated into society through policies that aim to appropriately balance costs and benefits.[55] The third is that the best approach to regulating surveillance is through tech-neutral legislation that applies to all surveillance technologies and does not single out specific ones for unique treatment.[56]

For the reasons that we have provided, we believe that these presumptions do not apply here and conclude that, at a minimum, moratoriums are justified because the conditions for consentability for facial recognition technology have not been met. Furthermore, face surveillance of all kinds presents a panoply of harms, most notably corrosion of collective autonomy through the chill of increased surveillance and machines indulge the fatally flawed notion of perfect enforcement of the law. Neither consent nor procedural frameworks like warrant requirements are sufficient to address these harms. As such, we argue face surveillance should be banned. Regulating the government without also imposing restrictions on technology companies is insufficient, but a promising start because, at present, government agents pose the greatest threats.

As Clare Garvie rightly observes, mistakes with facial recognition technology can have deadly consequences.[57] This means they can trample an individual's right to be free from bodily harm, the highest of the individual autonomy rights in Kim's

---

54. *See* Adam Thierer, *The Great Facial Recognition Technopanic of 2019*, MERCATUS CTR. (May 17, 2019), https://www.mercatus.org/bridge/commentary/great-facial-recognition-technopanic-2019.

55. *See, e.g.,* James O'Neil, *How Facial Recognition Makes You Safer*, N.Y. TIMES (June 9, 2019), https://www.nytimes.com/2019/06/09/opinion/facial-recognition-police-new-york-city.html; *America is Turning Against Facial-Recognition Software*, ECONOMIST (May 23, 2019), https://www.economist.com/united-states/2019/05/23/america-is-turning-against-facial-recognition-software.

56. *See, e.g.,* Judith Donath, *You Are Entering an Ephemeral Bio-Allowed Data Capture Zone*, MEDIUM (July 23, 2018), https://medium.com/@judithd/you-are-entering-an-ephemeral-bio-allowed-data-capture-zone-5ecafd2dbdaf.

57. Clare Garvie, *Facial Recognition Threatens Our Fundamental Rights*, WASH. POST (July 19, 2018), https://www.washingtonpost.com/opinions/facial-recognition-threatens-our-fundamental-rights/2018/07/19/a102703a-8b64-11e8-8b20-60521f27434e_story.html.

framework.[58]

> What happens if a system like this gets it wrong? A mistake by a video-based surveillance system may mean an innocent person is followed, investigated, and maybe even arrested and charged for a crime he or she didn't commit. A mistake by a face-scanning surveillance system on a body camera could be lethal. An officer alerted to a potential threat to public safety or to himself, must, in an instant, decide whether to draw his weapon. A false alert places an innocent person in those crosshairs.[59]

Lawmakers could regulate facial recognition a few different ways, and all but one will lead to an irrevocable erosion of obscurity and collective autonomy. When considering how to regulate private commercial use of facial recognition, lawmakers will be tempted to go back to that old standby regulatory mechanism that they always reach for when they lack political capital, resources, or imagination: consent. Consent is attractive because it pays lip service to the idea that people have diverse preferences, it's steeped in the law, and at a glance appears to be a compromise between competing values and interests. But as Kim demonstrated and we argue, it is fool's gold for facial recognition technologies, especially face surveillance. Even highly regulated and constrained use of facial recognition technology that has been agreed to will lead to an erosion of obscurity and a harm to our collective autonomy without actually serving our individual autonomy interests.

The problem is that there aren't many proven alternatives to consent regimes for commercial use of facial recognition that go beyond mere procedural frameworks. If the E.U.'s General Data Protection Regulation is any guide, the most prominent alternative to legitimize collection and processing of face biometric data is to require companies to have a "legitimate interest" in doing so.[60] But

---

58. NANCY KIM, CONSENTABILITY: CONSENT AND ITS LIMITS (2019).

59. Garvie, *supra* note 56.

60. General Data Protection Regulation 2016/679 of May 25, 2018, Lawfulness of Processing, art. 6(1)(f), http://www.privacy-regulation.eu/en/article-6-lawfulness-of-processing-GDPR.htm; *Recommendations for Implementing Transparency, Consent, and Legitimate Interest Under the GDPR*, CTR. FOR INFO. POL'Y LEADERSHIP (May 17, 2017), https://www.huntonprivacyblog.com/wp-content/uploads/sites/28/2017/06/cipl_recommendations_on_transparency_consent_and_legitimate_interest_under_the_gdpr_-19_may_2017-c.pdf ("Legitimate interest may be the most accountable ground for processing in many contexts, as it requires an assessment and balancing of the risks and benefits of processing for organisations, individuals[,] and society . . . . The legitimate interests to be considered may include

what constitutes a "legitimate interest" is notoriously slippery and subject to drift. Lawmakers have yet to get serious in using this concept to significantly rein in the power wielded by data controllers.

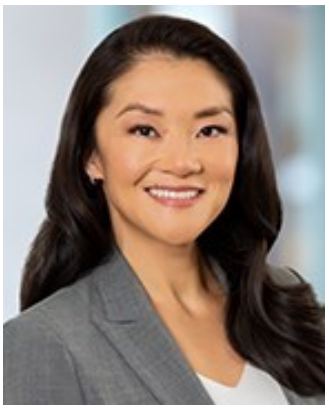
So, if facial recognition becomes entrenched in the private sector by procedural frameworks, that means that in addition to a warrant framework's accretion problem, the government will also have a backdoor to retroactive surveillance via the personal data industrial complex. Through public/private cooperation, surveillance infrastructure will continue to be built, chill will still occur, harms will still happen, norms will still change, collective autonomy still will suffer, and people's individual and collective obscurity will bit by bit continue to diminish.

The end result is that even if advocates of consent and warrant requirements got everything on their wish list, society would still end up worse off. We would suffer unacceptable harm to our obscurity and collective autonomy through a barrage of I agree buttons and search warrants powered by government and industry's unquenchable thirst for more access to our lives. There is only one way to stop the harms of face surveillance. Ban it.

---

the interests of the controller, other controller(s), groups of individuals[,] and society as a whole."); *CIPL Examples of Legitimate Interest Grounds for Processing of Personal Data*, CTR. FOR INFO. POL'Y LEADERSHIP (Mar. 16, 2017), https://iapp.org/media/pdf/resource_center/final_cipl_examples_of_legitimate_interest_grounds_for_processing_of_personal_data_16_march_2017.pdf.

Crowell



Jennie Wang VonCannon, CIPP/US

Partner

Los Angeles
jvoncannon@crowell.com
Phone: +1.213.310.7984

Practices

- White Collar and Regulatory Enforcement
- Privacy and Cybersecurity
- Litigation & Trial

Admissions

- California
- U.S. Supreme Court
- U.S. Court of Appeals, Ninth Circuit
- U.S. District Court, Central District of California
- U.S. District Court, Northern District of California

Jennie VonCannon is a trial lawyer with a proven track record of success in both the courtroom and the boardroom — with extensive experience in white collar defense and cybersecurity matters. Jennie helps clients in crisis with internal investigations, law enforcement and regulatory inquiries and subpoenas, and cybersecurity and privacy incidents. Her impeccable judgment has been honed over 11 years as a federal prosecutor, the last three of which she served with distinction as the deputy chief of the Cyber and Intellectual Property Crimes Section of the National Security Division of the U.S. Attorney's Office for the Central District of California.

Jennie is a Certified Information Privacy Professional, the chair and founding member of the Los Angeles County Bar Association Privacy and Cybersecurity Section, and an executive committee and life member of the Women Lawyers Association of Los Angeles. She has spoken at numerous industry programs on key topics such as the California Privacy Rights Act, ethical duties of technology competence, and cybersecurity.

Admissions/Affiliations

Professional Activities and Memberships

- Secretary and Life Member, Women Lawyers Association of Los Angeles (WLALA)
- Chair and Founding Member, Los Angeles County Bar Association (LACBA) Privacy and Cybersecurity Section
- Former Board Member, Southern California Chinese Lawyers Association (SCCLA)
- Member, Women's White Collar Defense Association (WWCDA)

Crowell

Representative Matters

- Secured across-the-board acquittals for a doctor charged with 33 felony counts of federal health care fraud after she used her forensic analysis experience to put on a highly technical defense case.
- On behalf of the boards of directors, investigated multiple instances of allegations of sexual assault or misconduct against the founders of high-profile companies facing intense media scrutiny.
- Has counseled companies experiencing ransomware attacks and data breach incidents regarding incident response and privacy notifications, working seamlessly with forensic firms and liaising with law enforcement as well as providing guidance to prevent such incidents.

Education

- University of California, Berkeley, B.A. English & Political Science (2001) with Honors
- University of California, Berkeley School of Law, J.D. (2004)

Speaking Engagements

- "SEC Rules on Cybersecurity," Society for Corporate Governance (December 2022). Speaker: Jennie VonCannon.
- "Ninth Annual Legal Careers in Cybersecurity, Privacy & Information Law," ABA Science & Technology Section (November 2022). Speaker: Jennie VonCannon.
- “Pieces of the Puzzle: Using Direct and Cross to Frame Your Narrative for Closing,” American Bar Association’s Professional Success Summit MCLE Skills Program (October 2022). Speaker: Jennie VonCannon.
- "Ethics Pitfalls and How to Avoid Them," Southern California Chinese Lawyers Association’s Phoenix Rising MCLE Program (January 2022). Speaker: Jennie VonCannon.
- "Security Risks in Cyber Space and Amended Rule 1.1," Consumer Attorneys Association of Los Angeles Convention (September 2021). Speaker: Jennie VonCannon.
- "Understanding the California Privacy Rights Act (CPRA)," National Asian Pacific American Bar Association Data Security and Privacy Committee Webinar (April 2021). Moderator: Jennie VonCannon.
- "The Impact of *Carpenter* and *Facebook*: What Can the Government Legally Access and How, and What Do Private Companies Have to Disclose in Response to Subpoenas?," Women Lawyers Association of Los Angeles and Los Angeles County Bar Association MCLE Program (December 2020). Speaker: Jennie VonCannon.
- "The Ethical Duty of Technology Competence," Women Lawyers Association of Los Angeles MCLE Program (August 2020). Speaker: Jennie VonCannon.

- "The Impact of the 2020 U.S. Cyberspace Solarium Commission Report's Findings on the Los Angeles Business and Legal Communities," Los Angeles County Bar Association (June 2020). Speaker: Jennie VonCannon.
- "Cryptocurrency and Cybersecurity," U.S. Securities and Exchange Commission Joint Regional Conference (June 2019). Speaker: Jennie VonCannon.
- "Strategies for IP Protection in China," U.S. Patent and Trademark Office China IP Road Show (June 2019). Speaker: Jennie VonCannon.
- "Cybersecurity in the Digital Age," Professional Fiduciary Association of California 24th Annual Educational Conference (May 2019). Speaker: Jennie VonCannon.
- "The Recent Surge in U.S. Law Enforcement against Chinese Companies," Cornell Law School Alumni Association MCLE Program (April 2019). Speaker: Jennie VonCannon.

Publications

- "Uber Scrutiny Of Cybersecurity," *The Daily Journal- subscription required* (January 10, 2023). Author: Jennie Wang VonCannon, CIPP/US.

Press Coverage

- Midterm Machinations: Law Firms Prepare For The November Congressional Reshuffle
  October 28, 2022 — The National Law Journal

- Cyber Lawyers Consider Impact Of Uber Executive's Conviction
  October 10, 2022 — Daily Journal (subscription required)

- Crowell Snags Former Federal Prosecutor In California
  October 6, 2022 — The Recorder

- Ex-Prosecutor Joins Crowell & Moring Cybersecurity Team
  October 6, 2022 — Law360

- Crowell Looks To Bolster Cyber Group As Feds Respond To Threats
  October 6, 2022 — Bloomberg Law

Press Releases

- Cyber Trial Lawyer and Former Assistant U.S. Attorney Jennie Wang VonCannon Joins Crowell & Moring (October 6, 2022)