Modern solutions to modern problems: leveraging the latest in data privacy and algorithmic fairness

Rachel Cummings, Ph.D.
November 21, 2022

This note[1] is written in response to the Federal Trade Commission's request for public comment, "Commercial Surveillance ANPR, R111004," on new rules to govern "the ways in which companies collect, aggregate, protect, use, analyze, and retain consumer data, as well as [how parties] monetize that data in ways that are unfair or deceptive." These comments focus on presenting research on privacy enhancing technologies (PETs) alongside transparency and fairness in the context of algorithmic inference. This note is structured in the following sequence:

- Current privacy and algorithmic practices highlight a regulatory imperative
  - "Notice and consent" is inadequate for protecting consumer privacy [Q73, Q80]
  - Gaps exist in current approaches to privacy protection [Q10, Q11]
  - Algorithmic decision-making is subject to common pitfalls [Q53, Q55]
- Rulemaking can build on existing best practices with new technologies
  - Employing state-of-the-art privacy-related technology can better protect consumers [Q48, Q83, Q11]
  - Algorithmic pitfalls can be avoided by explicit and context-dependent accounting for error, bias, and fairness [Q53, Q54, Q67]
  - Explaining algorithmic decisions arms consumers with understanding and empowers them to challenge automatic decision-making systems [Q89]
  - Explaining decision-making processes allows for more effective disclosures to at-risk populations [Q58, Q90]

Rachel Cummings is an Assistant Professor in the Department of Industrial Engineering and Operations Research at Columbia University. She is also an Affiliate Professor in the Department of Computer Science and a Member of the Data Science Institute at Columbia University. Her research focuses on data privacy and ethical AI systems, and her work has addressed topics such as the design of privacy-preserving machine learning systems, the relationship between fairness and privacy, explainability of algorithmic decision systems, economic impacts of privacy policy, legal interpretations of PETs such as differential privacy, economic incentives surrounding the use of consumer data, and human-centric explanations of PETs. She serves on the Advisory Board for the Future of Privacy Forum, on the Association for Computing Machinery's US Technology Policy Committee in the Privacy Subcommittee, and on

[1] The author thanks Stephen Chan, Nicholas Goutermout, James Palano, Carley Reardon, and Bartley Tablante for research assistance.

the Institute of Electrical and Electronics Engineers' Standards Association in the Working Groups for Algorithmic Bias and Transparent Data Governance.

Current privacy and algorithmic practices highlight a regulatory imperative

Further delaying the introduction of checks related to data and algorithmic decision-making enables increasing harms to the public, as pernicious practices can both accumulate and become entrenched as norms.

"Notice and consent" is inadequate for protecting consumer privacy

Relevant questions: 73 (effectiveness of consent), 80 (effectiveness of opt-out choices)

Informed user consent has been considered a justified source of carte blanche for technology firms to do anything with user data within their written policies. This paradigm is problematic, as the very notion of "informed consent" is deeply misleading. As the Commission has noted in its ANPR, there is considerable evidence that consent for privacy practices can never be fully informed and that the implications of data practices are highly complex and are not understood by consumers.[2] Even in the case of well-informed consumers who choose to engage with a privacy policy, research has found it is nearly impossible for the consumer to accurately assess the risks and trade-offs associated with the consent choice.[3] Most consumers do not even attempt this level of engagement with the risks, with 85% spending less than 10 seconds reading Google's privacy policy according to one study.[4] Applications of ML models at scale are instances in which techniques and usage of consumer data are functionally impossible for the average consumer to understand. Moreover, technology companies typically design the user interface of consent prompts as so-called "dark patterns," which are suggestive – and often deceptive – and designed to encourage consumers to click "agree" faster and agree to provide more data.[5] Recently, Google paid a $392 million settlement after it was charged with continuing to collect geolocation data on consumers who turned off a "Location History" setting in their

[2] Federal Trade Commission. Trade Regulation Rule on Commercial Surveillance and Data Security. p. 51275. August 22, 2022. Available at https://www.federalregister.gov/d/2022-17752/p-31

[3] Alessandro Acquisiti and Jens Grossklag. What Can Behavioral Economics Teach Us about Privacy? *Digital Privacy: Theory, Technologies and Practices*, pp. 6-7. 2007. Available at https://www.heinz.cmu.edu/~acquisti/papers/Acquisti-Grossklags-Chapter-Etrics.pdf

[4] UK Competition and Markets Authority. Online Platforms and Digital Advertising – Market Study Final Report. 2020. Available at https://assets.publishing.service.gov.uk/media/5efc57ed3a6f4023d242ed56/Final_report_1_July_2020_.pdf

[5] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. 2020. Available at https://doi.org/10.1145/3313831.3376321

account settings[6,7] Finally, consent is ineffective in instances in which opting-out is not a reasonable option. Currently, major technology companies have removed opting out as a valid option for consumers since refusing the terms of service of these companies is akin to opting out of internet services.. It is imperative that "informed consent" no longer be used as justification to neglect the necessary practices described herein.

Gaps exist in current approaches to privacy protection

Relevant questions: 10 (kinds of data that should be subject to new regulation), 11 (checks companies rely on to not cause harm to consumers)

Current and proposed legislation and regulation related to data privacy have the right intentions but are inadequate to protect user privacy from even well-known re-identification techniques. Much of this legislation deems certain kinds of data worthy of special protection, leaving the rest as fair game for any data practice. For instance, the current approach in legislation of distinguishing between Personal Identifiable Information (PII) and non-PII, and specifying different treatment for each, is not sufficient to ensure data protection. This is because using non-PII can still often render individuals as identifiable when used in combination with other data. In one famous example, publicly available information on IMDb was used to de-anonymize users in the Netflix Prize dataset.[8] The risk of re-identification also extends into domains with even stronger expectations of privacy; for example, publicly-available voter registration information was sufficient to identify individual health records in an anonymized dataset released by the Massachusetts Group Insurance Commission.[9] The same lack of protection has been found to hold true for distinctions like "sensitive data" (i.e., other features may be correlated or related to sensitive attributes) and "protected categories" because it is impossible to define all possible proxies or combinations of other features that may together serve as a proxy. As such, while compliance may be relatively straightforward with rules requiring special treatment for different data categories like PII, compliance with these rules is not sufficient to ensure consumer protection.

Aggregating data also does not necessarily eliminate privacy risk. Aggregated data concerns are especially prevalent in contexts involving ML and large datasets. Despite the fact that the mechanics of most ML algorithms involve some degree of aggregation and/or anonymization of the input data, many common ML algorithms can memorize individual data entries during

---

[6] Cecilia Kang. Google Agrees to $392 Million Privacy Settlement With 40 States. *New York Times*. November 14, 2022. Available at https://www.nytimes.com/2022/11/14/technology/google-privacy-settlement.html

[7] Ryan Nakashima. AP Exclusive: Google tracks your movements, like it or not. *Associated Press*, 13 August 2018. Available at https://apnews.com/article/north-america-science-technology-business-ap-top-news-828aefab64d4411bac257a07c1af0ecb

[8] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy*, pp. 111-125. 2008. Available at https://ieeexplore.ieee.org/document/4531148

[9] Latanya Sweeney. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, *Data Privacy Working Paper* 3. Pittsburgh. 2000.

training and reproduce these entries as a part of their output. For example, deep learning algorithms for word prediction have leaked Social Security Numbers and credit card numbers when trained on a corpus that included such data.[10] Additionally, techniques used in many recommender systems can leak information across users through their personalized recommendations.[11] These systems are not sufficiently anonymized and hence fail to meet meaningful legal standards of privacy like GDPR compliance.[12]

Furthermore, even firms with sophisticated approaches to privacy preservation have little incentive to be transparent about their methods. For example, in 2017, Apple's choice of parameters in their implementation of differential privacy was alleged to provide insufficient privacy protections to users.[13] Without effective disclosure of key parameters (e.g., epsilon for differential privacy, as discussed below), consumers and regulators alike do not have enough information to determine how effective any privacy protections these technologies provide are.

Algorithmic decision-making is subject to common pitfalls

Relevant questions: 55 (weight given to automated decision-making system outputs), 53 (inevitability of algorithmic error)

Automated systems make consequential decisions, with some of these decisions having significant impacts on our lives, including literal life or death consequences.[14] Algorithms are pervasive in decisions about whom to hire,[15] school assignment for children,[16] determination of

[10] Nicholas Carlini, Chang Liu, Jernej Kos, Ulfar Erlingsson, and Dawn Song. The Secret Sharer: Evaluating and testing unintended memorization in neural networks. *Proceedings of the 28th USENIX Security Symposium, USENIX Security '19*, 267–284. 2019.
[11] Joseph A. Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten, and Vitaly Shmatikov. "You Might Also Like:" Privacy Risks of Collaborative Filtering. *2011 IEEE Symposium on Security and Privacy,* 231–246. 2011.
[12] Rachel Cummings and Deven Desai. The Role of Differential Privacy in GDPR Compliance. *Workshop on Responsible Recommendation*. 2018.
[13] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in Apple's implementation of Differential Privacy on macOS 10.12. 2017. Available at https://arxiv.org/pdf/1709.02753.pdf
[14] For example, autonomous vehicles (e.g., Sven Nyholm and Jilles Smids. The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?. *Ethical Theory and Moral Practice* 19, 1275–1289. 2016. Available at https://doi.org/10.1007/s10677-016-9745-2).
[15] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* October 10, 2018. Available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[16]Sidney Fussell. New York City Wants to Audit the Powerful Algorithms That Control Our Lives. *Gizmodo*. December 14, 2017. Available at https://gizmodo.com/new-york-city-wants-to-audit-the-powerful-algorithms-th-1821305284

jail time,[17] and even healthcare coverage.[18] The weight of these automated decisions is troubling in instances when the decision-making systems make errors, are systematically biased, and use poorly-understood reasoning.

Algorithmic decisions will always involve a degree of error. When drawing conclusions about a large population from a finite sample of observations, error is unavoidable. Additional data can only reduce but never eliminate error.[19] Like error, bias is also inescapable when building statistical models from data. Bias can be introduced into models in a variety of forms, such as biased training data, bias in the algorithmic decision-making process, or unforeseen bias in downstream decisions. Models must be carefully constructed on datasets that both accurately reflect the population for the task being studied and do not encode existing or historical biases.[20] Instances of discrimination have occurred when algorithms designed for hiring and facial recognition have been trained on biased datasets.[21,22,23] Bias may still arise even when algorithms are trained on unbiased datasets due to discrimination via proxy.

Rulemaking can build on existing best practices with new technologies

New rules and regulations should build on the strong foundation of existing best practices related to privacy, data use, and algorithmic decision-making, and should recognize significant advances in technology to enhance privacy and the auditability of algorithms.

---

[17]For an algorithm used in sentencing, see (Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. May 23, 2016. Available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm) and (William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe. 2016. Available at https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf)

[18] Colin Lecher. What happens when an algorithm cuts your health care. *The Verge*. March 21, 2018. Available at https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[19] Vladimir Vapnik. Principles of risk minimization for learning theory. *Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS'91)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 831–838. 1991. Available at https://proceedings.neurips.cc/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf

[20] For examples of problems that may arise when databases are not representative, see Appendix A of Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. 2012. Available at https://doi.org/10.1145/2090236.2090255

[21] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge and A. K. Jain, "Face Recognition Performance: Role of Demographic Information," in *IEEE Transactions on Information Forensics and Security*. Vol. 7, no. 6, 1789–1801. December 2012. Available at https://doi.org/10.1109/TIFS.2012.2214212

[22] Erin Winick. “All automated hiring software is prone to bias by default,” *MIT Technology Review*. December 23, 2018. Available at: https://www.technologyreview.com/2018/12/13/138670/all-automated-hiring-software-is-prone-to-bias-by-default/

[23] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in *Proceedings of Machine Learning Research*. Vol. 81, 77–91. 2018. Available at https://proceedings.mlr.press/v81/buolamwini18a.html

Employing state-of-the-art privacy related technology can better protect consumers

Relevant questions: 48 (data minimization and purpose limitations and algorithmic learning), 83 (information about commercial surveillance companies should be required to disclose), 11 (checks companies rely on to not cause harm to consumers)

Rulemaking would do well to include modern privacy enhancing technologies (PETs) alongside more traditional notions of data minimization and purpose limitation to better preserve privacy and minimize the exposure of consumer data. Differential privacy is one modern PET that can ensure privacy by adding controlled amounts of noise to data analysis processes to hide the impact of one person's data. Differential privacy guarantees that no adversary can infer any individual's data from the output of the algorithm, regardless of the adversary's computational power or outside information. These guarantees are governed by a privacy parameter, epsilon: smaller epsilon values mean stronger privacy guarantees (i.e., it is harder for adversaries to infer any individual's data). Epsilon also composes smoothly across multiple uses of a dataset, enabling the same dataset to be used in several analyses without violating the privacy guarantees. This privacy parameter allows the practitioner to transparently and concretely communicate the level of privacy provided by the system. Past work has empirically shown that training ML models with differential privacy has been proven to protect against membership inference attacks, model inversion attacks, reconstruction attacks, reidentification attacks, and predicate singling out attacks.[24,25,26] Additionally, differentially private algorithms have already

[24] Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. *Proceedings of the 28th USENIX Security Symposium, USENIX Security '19*. 267–284. 2019.
[25] Aloni Cohen and Kobbi Nissim. Towards formalizing the GDPR's notion of singling out. *PNAS* Vol. 117, no. 15, 8344–8352. March 31, 2020. https://www.pnas.org/doi/full/10.1073/pnas.1914598117
[26] Simson Garfinkel, John M. Abowd, and Christian Martindale. Understanding Database Reconstruction Attacks on Public Data. *ACM Queue*. Vol. 16, issue 5. November 28, 2018. Available at: https://queue.acm.org/detail.cfm?id=3295691

been deployed at large scale by organizations such as Apple,[27,28,29] Google,[30] Microsoft,[31] Uber,[32] and the U.S. Census Bureau.[33,34]

Firms must also be transparent about data privacy practices. As an example, businesses may claim to preserve privacy using differential privacy, but these claims can be meaningless or misleading if not accompanied by disclosure of other key technical details, such as the value of the privacy parameter epsilon[35] that was used. Additionally, differential privacy can be implemented either in the central model, where individuals provide data directly to a trusted curator, or the local model, where individuals add noise to their own data before sharing it with an untrusted curator. Specifying this trust model is critical for users to understand how their data can be viewed, accessed, and analyzed.[36,37] Finally, there exist many differentially private algorithms, all of which are governed by the same epsilon parameterization, but have different distributional implications for the outcomes, which may be relevant for consumers. Beyond

---

[27] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection Against Reconstruction and Its Applications in Private Federated Learning. arXiv preprint 1812.00984. 2018. Available at https://arxiv.org/abs/1812.00984

[28] John Duchi and Ryan Rogers. Lower Bounds for Locally Private Estimation via Communication Complexity. *Proceedings of the Thirty-Second Conference on Learning Theory*, in *Proceedings of Machine Learning Research*. Vol. 99, 1161–1191. 2019. Available from https://proceedings.mlr.press/v99/duchi19a.html

[29] Rachel Cummings, Vitaly Feldman, Audra McMillan, and Kunal Talwar. Mean Estimation with User-level Privacy under Data Heterogeneity. Forthcoming at NeurIPS 2022.

[30] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. Association for Computing Machinery, New York, NY, USA, 1054–1067. 2014. Available at https://doi.org/10.1145/2660267.2660348

[31] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3574–3583. 2017.

[32] Noah Johnson, Joseph P. Near, and Dawn Song. Towards practical differential privacy for SQL queries. *Proc. VLDB Endow*. Vol. 11, issue 5, 526–539. January 2018. Available at https://doi.org/10.1145/3177732.3177733

[33] John M. Abowd. The U.S. Census Bureau Adopts Differential Privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2867. 2018. Available at https://doi.org/10.1145/3219819.3226070

[34] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets Practice on the Map. *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE '08)*. IEEE Computer Society, USA, 277–286. 2008. Available at https://doi.org/10.1109/ICDE.2008.4497436

[35] The parameter epsilon captures the amount of information leaked about an individual, and can range from zero to infinity, allowing the extremes of complete privacy and complete disclosure, and everything in between.

[36] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. "I need a better description": An Investigation Into User Expectations For Differential Privacy. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. Association for Computing Machinery, New York, NY, USA, 3037–3052. 2021. Available at https://doi.org/10.1145/3460120.3485252

[37] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. Improving Communication with End Users About Differential Privacy. Appeared at Theory and Practice of Differential Privacy, 2022.

simply disclosing these technical details, businesses should use explanatory methods to aid consumers in understanding them in layman's terms.

Future rulemaking for PETs should both encourage the effective and transparent use of techniques like differential privacy. Since there is currently very little guidance for practitioners about appropriate choices of parameters like epsilon to set meaningful privacy guarantees,[38] such rules should also provide context-based guidance for appropriate choice of epsilon in order to be effective.

Algorithmic pitfalls can be avoided by explicit and context-dependent accounting for error, bias, and fairness

Relevant questions: 53 (costs and benefits of automated decision-making systems), 54 (best ways to measure algorithmic error), 66 (how to measure algorithmic discrimination), 67 (how algorithmic discrimination should be addressed)

When guided by proper fairness principles and accounting for bias and error, the use of automated decision-making can illuminate biases and mistakes that are swept under the rug in human judgment. Algorithms' decisions can be audited: they are replicable and can be analyzed to understand how changes in the input would or would not have led to different outcomes. They can also be guided by consistent measurable approaches to fairness and provide explicit estimates of confidence and error. By harnessing these benefits of algorithmic systems, practitioners can ensure the greatest positive impact of automated decision-making.

It is important to have a good understanding of the error of a given model to know how accurate the model's prediction will be when applied to the real world. For instance, certain binary-outcome algorithms can provide a confidence score for a given prediction (e.g., "yes with 83% certainty"). Beyond being necessary, the choice of approach in estimating model error must be specific to the context in which it is used. Sometimes this choice is more well-defined, as in choosing how to formalize accuracy for a classifier (e.g., classification accuracy, false negative rate). Other techniques, like synthetic data generation, do not have a single well-defined notion of accuracy. Instead, these approaches have multiple reasonable accuracy metrics, which crucially requires the use of context-specific knowledge in choosing an accuracy notion. Practitioners must carefully consider and choose which method of measuring error is appropriate for the task.

Creators of algorithmic decision-making systems must be cautious in their deployment and sensitive to the negative impacts that can result from bias and discrimination.[39] There are a

[38] Dwork, Cynthia, Nitin Kohli, and Deirdre Mulligan. "Differential Privacy in Practice: Expose Your Epsilons!". *Journal of Privacy and Confidentiality*. Vol. 9, no. 2. 2019. Available at https://doi.org/10.29012/jpc.689

[39] For example, Microsoft recently stopped providing tools that enabled prediction of age, gender, or emotions. (Kashmir Hill. Microsoft Plans to Eliminate Face Analysis Tools in Push for "Responsible A.I." *New York Times*. June 21, 2022. Available at https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html)

range of solutions for auditing the degree of bias in training data and resulting algorithms.[40] A type of bias, discrimination via proxy is particularly challenging to avoid because one cannot foresee all possible features that may be correlated with protected attributes. One solution to this problem is multicalibration,[41] which is a technique designed to simultaneously provide fair treatment across multiple overlapping categories. Unlike approaches that require explicitly specifying all possible protected categories or combinations thereof, multicalibration automatically protects all identifiable sub-groups within a dataset.

In addition to choosing approaches for handling error and bias, the design and specification of algorithms also involves difficult trade-offs between notions of fairness that must be informed by modern frameworks and how the model's predictions are translated into decisions. Not only is there no single universally correct definition of fairness,[42] but reasonable definitions of fairness can be incompatible with one another. For instance, classification parity and calibration, two reasonable definitions of fairness, have been shown mathematically to be mutually exclusive.[43] As an example of mutually exclusive definitions of fairness, consider a model used to determine whether a person has some risk factor, and suppose that a greater proportion of low income individuals have this risk factor. Then, any model for estimating the probability that someone has this risk factor must either have different classification accuracy for low and high income individuals or have different false positive rates for low and high income individuals. Determining which measure of fairness is more appropriate is not a question of choosing a superior technical methodology but one of choosing the policy best suited to the application (e.g., if the consequences of being identified as having the risk factor could have substantial negative impacts, then equal false positive rates might be more appropriate, but if accurate probabilistic predictions are important, then classification accuracy might be preferable). Practitioners must deliberately choose to be guided by a framework for fairness to ensure that consequential choices in algorithmic design are made explicitly and deliberately.

Explaining algorithmic decisions arms consumers with understanding and empowers them to challenge automatic decision-making systems

Relevant question: 89 (explaining implementation and use of automated decision-making systems in reaching decisions)

Significant progress is being made to illuminate "black box" algorithms. For instance, explainer models have been developed, which are simpler, interpretable, easier to understand models designed to approximate complex algorithms over various points in the input data. These

[40] Many popular techniques can be found at, for example, https://fairlearn.org/
[41] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. *Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research* 80:1939-1948. 2018. Available at https://proceedings.mlr.press/v80/hebert-johnson18a.html.
[42] Arvind Narayanan. Tutorial: 21 fairness definitions and their politics. March 1, 2018. Available at: https://youtu.be/jIXIuYdnyyk
[43] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)*. 2017. Available at https://arxiv.org/abs/1609.05807

explainer models can be global[44,45] (intended for use on all data), local[46] (intended for use only on a specialized subpopulation), or ensembles of local explainers[47] that provide different explanations for different subpopulations. With all of these methods, there is an inherent tradeoff between the fidelity of the explainer (i.e., whether the prediction matches that of the black box model), coverage (how broadly can the explainer be applied), and interpretability (more complex models may produce better predictions but be less understandable themselves).

As increasingly sophisticated algorithmic decision-making techniques are developed, these explanatory methods are increasingly important for imparting insight and recourse to those who are subject to these decisions and analyses. For example, someone denied a loan might be told the main reasons were a high debt to income ratio and short credit history. Some of this is actionable (e.g., make future credit card payments on time) and some is not (e.g., have a 10-year-longer credit history). This disclosure is important both because decisions might have been based on incorrect information (e.g., the wrong credit history was used) and because some non-actionable criteria are inappropriate for decision making (e.g., race).

Explaining decision-making processes allows for more effective disclosures to at-risk populations

Relevant questions: 90 (comprehensible disclosures), 58 (protecting non-English speaking communities from abusive data practices)

Equally important as explaining the decision logic of algorithms is explaining that information in a way that can be understood by those who are affected by automated decisions. Approaches to explaining algorithmic techniques like machine learning and differential privacy can be used to promote effective disclosure, particularly for communities at risk for fraud and abuse like non-English speakers. Similar techniques that leverage visual explanations and low-reading-comprehension-level explanations (e.g., “privacy nutrition labels”[48]) can be useful both for these groups, and more generally for everyday consumers, who likely lack the expertise

[44] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable Decision Sets: A Joint Framework for Description and Prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. 2016. Available at: https://doi.org/10.1145/2939672.2939874
[45] Hamsa Bastani, Osbert Bastani, and Carolyn Kim. Interpreting predictive models for human-in-the-loop analytics. FATML Workshop. 2017. Available at https://hamsabastani.github.io/interp.pdf
[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. 2016. Available at https://doi.org/10.1145/2939672.2939778
[47] Qiaomei Li, Rachel Cummings, and Yonatan Mintz. Optimal Local Explainer Aggregation for Interpretable Prediction. arXiv preprint 2003.09466. 2022. Available at https://arxiv.org/abs/2003.09466
[48] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A “nutrition label” for privacy. *Proceedings of the Fifth Symposium On Usable Privacy and Security (SOUPS)*. ACM, 1–12. 2009.

to interpret disclosures themselves.[49] These approaches should be shaped to fit the techniques, use cases, contexts, and users to maximize effectiveness.[50]

[49] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. Improving Communication with End Users About Differential Privacy. Appeared at Theory and Practice of Differential Privacy. 2022.

[50] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2021. "I need a better description": An Investigation Into User Expectations For Differential Privacy. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21). Association for Computing Machinery, New York, NY, USA, 3037–3052. https://doi.org/10.1145/3460120.3485252