Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems

Nitin Kohli[*] NITIN.KOHLI@BERKELEY.EDU *School of Information, UC Berkeley*

Renata Barreto RBARRETO@BERKELEY.EDU *Boalt Hall School of Law, UC Berkeley*

Joshua A. Kroll KROLL@BERKELEY.EDU *School of Information, UC Berkeley*

1. Introduction

Words such as fairness, accountability, transparency, bias, ”explanation” and many others suffer from linguistic conflation. Scholars across different fields study disparate topics using distinct approaches, outlooks, and methods, while using shared terminology to describe distinct ideas. Even seemingly straightforward terms such as ”algorithm”, ”artificial intelligence”, and ”machine learning” have murky boundaries and contested histories. Research at the interface of software systems and their human context (as well as practical policymaking and specifically drafting and interpreting the law) necessarily engages concepts across disciplines. However, because scholars and practitioners in different disciplines use the same words to mean different things, it can be challenging to advance understanding in a way that affects research or practice in different communities. Instead, faced with the question of how to describe concepts precisely, scholars and practitioners often double down on their existing disciplinary preconceptions, believing that resorting to their particular approach to rigor will surely convince those of different backgrounds.

In an effort to build community around research in fairness, accountability and transparency, this tutorial presents the fruits of our research into the use of vocabulary by different stakeholder communities. One approach to facilitating better communication when terms are not available that cross cultural boundaries is to describe concepts. Covering core technical ideas in machine learning and data science for a broad audience, we present a set of core terms from our research and describe example scenarios in which an algorithmic system impacts the world in order to elucidate how these systems are created and used, how they function, and how they are described and evaluated by different communities. Further, we examine how different groups understand, forecast, and approach the issues that may come up in the design, development, deployment, and evaluation of algorithmic systems in an effort to educate the tutorial audience about where to look for gaps in disciplinary understanding. Our focus in choosing example scenarios is on situations where terminological conflation leads to misunderstanding and confusion. In describing the use of these terms by different communities, we also describe how those communities understand contrast concepts for those terms (that is, what concepts each community thinks are opposed to the concept that community means by the terms and why). We also consider whether disagreement about the meaning or scope of one term leads to downstream misunderstandings about other terms and about the propriety or applicability of particular interventions.

Specifically, we consider in this tutorial the triad of fairness, accountability, and transparency, which gives the conference its name. We also consider a fourth concept, which has seen much salience across fields and uptake in the world of public policymaking, namely the concept of explanations for automated decisions as well as interpretability and intelligibility for data-driven models (sometimes referred to by the moniker ”xAI”). By choosing these terms out of our lexicon, we do not aim to frame the entire research conversation in terms of them or suggest that they are the most important terms. Rather, we aim to elucidate disciplinary differences in understanding in order to help the community

[*] All three authors contributed equally to this work

understand what frames are available and what makes them useful.

Our discussion of terminology provides the opportunity to consider the frontier of research on the core topics represented by the terms in our lexicon. For example, we discuss the forefront of research in machine learning fairness, viewed through the lens of how the term fairness is used by computer scientists, practicing lawyers, law scholars, social scientists, philosophers, and others. This discussion allows us to suggest approaches to bridging the gaps between these constructs and highlights opportunities for each community to make use of the work being done by outside scholars and practitioners.

2. Fairness

Fairness, along with bias and discrimination, guides the discussion on applications of machine learning, especially in high stakes decision-making. However, there is ongoing debate regarding the conceptualization and operationalization of fairness. Broadly, there exist four clusters of fairness definitions: philosophical, legal, social scientific, and technical (or computer science based criteria).

The scholarship on algorithmic fairness draws heavily on Rawls seminal text, A Theory of Justice. Rawls is especially useful in understanding algorithmic harms which are not captured by legal doctrine, such as loss of dignity (Hoffman, 2017). For example, how do different definitions of fairness speak to the discrepancy in internet search results between black girls and white girls, in which the former leads to an egregious amount of porn hits? (Noble, 2018). Political philosophy has long wrestled with notions of fairness, but under a different umbrella: egalitarianism, which deals with both the treatment of individuals and the distribution of valuable resources and varies tremendously by the strand, or version, of egalitarianism in question (Binns, 2018). Notably, in philosophy the term fairness is often used interchangeably with equity and equality, as well (see Dworkin (1981); Scanlon (2004); Anderson (1999)) . The roots of fairness in the Western philosophical canon are rhizomatic and wide-spread, so its critical to address these competing interpretations.

American jurisprudence acknowledges its origins in political philosophy; despite its resemblance to moral philosophy, legal doctrine has developed its own set of criteria to evaluate fairness in practice. The literature on machine learning overcorrects for the problem of bias by focusing on two interrelated legal concepts disparate treatment and disparate impact (Barocas and Selbst (2016); Lipton et al. (2017)). With foundations in labor law, disparate treatment argues that unequal behavior towards a protected class or individual is discriminatory and, therefore, unlawful. On the other hand, disparate impact focuses on the outcome among individuals or groups in a protected class. To make this distinction more concrete - an employer, for example, can apply a hiring rule that is racially neutral, but results in an unwarranted exclusion of women, a protected class. Algorithmic discrimination can proceed in either direction. However, the global reach of algorithmic decision making should give pause to scholars who formalize fairness criteria solely on Western legal norms and definitions. In the UK, where the legal system mirrors that of the US, disparate impact is understood as indirect or institutional discrimination (Binns, 2018). In a similar vein, protected classes are outgrowths of particular histories of social exclusion, so how do these vary across national lines? The question of how other countries and cultures think through fairness in algorithmic contexts remains a largely unexplored research agenda.

The social sciences, whose theories and methods are historically situated, have developed more nuanced understandings of discrimination. For example, social scientists count with structural discrimination to explain social phenomeonon. The unit of analysis here is at a collective or social level; this makes it challenging to reach consensus with legal scholars, for example, who often rely on the individual as the unit of analysis. sychologists developed the idea of implicit bias, an unconscious, stereotypical associations that affect behaviors and actions (Greenwald and Krieger, 2006). This definition becomes problematic when it is liberally applied to algorithmic decision making. For example, in Levendowksis piece, “How Copyright Law Can Fix Artificial Intelligences Implicit Bias Problem” (Levendowski, 2018), she equates implicit bias, a well-documented cognitive phenomenon present in wetware, with algorithmic bias, which occurs in software systems that must make their preconceptions explicit, even if the source of those preconceptions may be implicit or hidden to the creators or users of those data. Data contains both implicit and explicit biases; however, this should not be conflated with the type of bias that occurs at the level of the algorithmic system itself. Implicit bias is ultimately a human limitation. (Dwork and Mulligan, 2013). Moreover, its notoriously difficult to change ones implicit biases on the other hand, with the right training data and attribute-sensitive algorithm, we can alter the associations that algorithms produce. However, it is important not to attribute all bias merely to data and none to the algorithms making use of data. Additionally, social scientists often engage with the idea of structural discrimination to explain social phenomena.

Technical definitions on fairness and bias are varied and manifold. They include, but are not limited to, accuracy equity (Angwin et al., 2016), conditional accuracy equity (Dieterich et al., 2016), equality of opportunity (Hardt et al., 2016), disparate mistreatment (Zafar et al., 2017), predictive bias and statistical discrimination (Chouldechova, 2017). These examples do not exclusively focus on the internal workings of the algorithms themselves; instead, they often attempt to model human bias and behavior as well and use this as a point of departure or commonality with algorithmic bias. For example, Chouldvecha et al 2018 demonstrates that caseworks in Allegheny County, PA use proxies like zip code and race to draw conclusions about the likelihood of child abuse. Scholars distinguish between fairness criteria that examine the decisions versus the risk scores (Corbett-Davies et al., 2017). Markedly, there is no algorithm that can satisfy all major fairness criteria simultaneously, such as calibration and balance for the positive and negative class (Kleinberg et al., 2016). Consensus exists about the relevant trade-offs, such as public safety and fairness, but the thresholds remain a point of contention. Indeed, formalizing intuitions into mathematical formula and subsequently code do not erase the inherently normative dimension of fairness, a point which most researchers appear to agree with.

3. Accountability

Accountability is fundamentally about the answerability of actors for outcomes. For computer systems, this answerability can come at a few levels of detail: answerability follows most simply from accounting, the creation and maintenance of detailed records of what outcomes occurred, which actors contributed to those outcomes, and how they contributed. More broadly, accountability can refer to the responsibility or ownership of those outcomes and the way in which accountings are viewed in light of social, political, legal, and moral norms. In the law, this notion of responsibility is often coupled with notions of liability and punishment for misdeeds with a focus on accountability as review, oversight, and enforcement. Stone, Jabbra and Dwivedi (1989) define seven key types of accountability, based on the type of entity to whom we demand answerability and the grounding for the answerability: moral, administrative, political, managerial, market, legal/judicial, constituency relation, and professional.

The literature on accountability in computer science defines the property narrowly, as a kind of *trace property* of software systems.[1] Specifically, accountability is often defined as the property that, after a system runs, there exists a record which describes what the system did (i.e., what outputs or observable behaviors it had), what caused the system to do those things (i.e., what inputs were provided), and what agent within the system took particular actions (Haeberlen et al., 2007). Accountability in this view is heavily dependent on how a system is specified (i.e., what its designers meant for it to do in a particular context). Some notions of accountability relate to specific, normative compliance with requirements designed to capture political accountability, such as procedural regularity (Kroll, 2015). Others contextualize the narrow view of accountability as a restriction of the broader concept of *verifiability*, or provable fidelity to a predetermined set of requirements (Küsters et al., 2010). While technical accountability focuses heavily on the keeping of records, it is not the case that these records must of necessity be disclosed to

[1] In the analysis of software, a trace property is a property of particular executions, or “traces, of a program, rather than a property which is true of the program abstractly.

the affected parties to be useful - so long as those affected by the outcomes of a system can trust in the integrity and relevance of the records, the records may be leveraged by trusted actors such as oversight entities to broaden the scope of answerability (Kroll (2015); Kroll et al. (2017)). Here, the focus is on "what happened?, "why?, and "who took that action?

Looking more broadly, we can view accountability through the lens of responsibility and answerability to a particular entity, where the responsibility and answerability are vested with a particular entity rather than just a presumption about the value of keeping records. This view is carefully laid out by Nissenbaum (1996), who argues that such responsibility is held back in computer systems by a number of factors related to the way these systems are designed, implemented, and fielded. Others, namely Kroll et al. (2017), focus on vesting responsibility with the creators of computer systems as a way to sidestep the need to fully articulate a specification for a piece of technology *a priori*, rather suggesting that, at least in some cases, contested norms can be better disambiguated *ex post* during review or oversight. Desai and Kroll (2018) attempt to disambiguate technical notions of accountability-as-recordkeeping from the entrenched idea of accountability under the law, with a survey of discussions of accountability in the legal literature. The focus in this broader view is on questions of "what agent or entity is accountable for a particular outcome? and "to whom is that agent or entity accountable?

A particularly salient part of this broader view sees accountability as the requisite property to ensure punishment for misdeeds and the enforcement of commitments, either those imposed under the law or those made voluntarily by actors who control computer systems. This view captures uses of the term as applied to mistakes, errors, and explicit malfeasance, and is used in law enforcement contexts, to explain and justify liability regimes, and when discussing the legal concept of torts. Nissenbaum (1996) takes care, however, to distinguish accountability from liability, as liability may not require moral fault, while responsibility (from a philosophical perspective) requires both causal agency and moral fault. The focus in this context is on questions such as "which agent or entity should be punished in light of a particular bad outcome?

4. Transparency

At a fundamental level, transparency is used to describe various notions of openness. Openness can be as simple as the disclosure of what a system does or how it works or as abstract and complicated as considering how to include constituencies in the design of systems and processes. In some disciplinary silos, transparency also includes characteristics of understandability - i.e. interpretability, intelligibility, and explainability.

At the simplest level, transparency can be succinctly described as the disclosure of system internals to look under the hood of a given technology (see, e.g., Pasquale (2015)). Such disclosures often focus little on choices that are viewed as external to the system: the design choices that went into system, approaches to experimental design, normalization of data, etc. This notion of transparency is common in computer science (Kroll et al., 2017), as computer scientists consider algorithms and software to be concepts that exist fundamentally inside a computer (to the extent that they are reified at all). This idea also appears in legal scholarship and analysis, given the law's presumption that automated processes are fundamentally comprehensible (Desai and Kroll, 2018).

At a more abstract level (and especially in law and public policy specifically), we see transparency broadly referring to the *design*, the *function*, and the *inputs and outputs* of a technical system. More specifically, design aspects of transparency refer to the the openness about the existence and scope of a system (e.g., the HEW Advisory Committee on Automated Personal Data Systems Report, the OECD privacy guidelines [see Solove and Schwartz (2014)], and European privacy law such as the Data Protection Directive and its replacement the General Data Protection Directive, GDPR [see https://www.eugdpr.org]). While transparency about the design of a system is concerned with whether or not such a system exists and its broad contours, functional transparency considers information about the methods or processes by which information can be obtained on how a system functions (and is accessible by, e.g., FOIA Brauneis and Goodman (2018)). Finally, legal notions of transparency often consider how to examine a system's inputs and outputs in specific to determine how they relate to each other. In this reckoning, transparency is not the end itself, but rather an approach to understanding. The goal is to understand the rationale for the outputs by way of pertinent information about the inputs. As an example of this flavor of transparency, consider the way that the Fair Credit Reporting Act and Equal Credit Opportunity Act consider transparency as an explanatory tool (see Hoofnagle (2013), Brill (2015)).

Within the social sciences and statistics, transparency often takes this one step further. In addition to technical disclosures about the structure or inputs and outputs of a system, transparency also requires making available experimental design and analysis choices (Miguel et al. (2014);Gelman (2017)). For example: How were data collected? How were data cleaned? Are data representative of the world? What methods were used in the analysis? What were the choices of thresholds and other hyperparameters? Thus, transparency in the social sciences is aligned with reproducibility, enabling the verification and the use of methods by others.

In some contexts, the term transparency is used interchangeably with interpretability, explainability, and intelligibility (e.g., Doshi-Velez et al. (2017); Shah and Kesan (2003)). That is, many authors assume or imply that transparency automatically leads to understanding, or that a requirement for understanding can be met by increased transparency. This need not be the case, however. For example, even if we imagine that Google could open source their entire code base, it is unreasonable to assume that any individual truly interprets it. Further, it may not be the case that details about a system capture sufficient context to engender sufficient understanding (Kroll et al., 2017).

These diverse uses of the term transparency have the ability to cause confusion amongst cross disciplinary groups, each of which are coming into conversations with their own disciplinary-specific uses of such a term.

5. Explanation/Interpretability/ Intelligibility

Explanation has received much attention as a possible solution to governance problems in software systems (Doshi-Velez et al., 2017), but different communities understand explanation and interpretability in substantially different ways.

To a machine learning researcher, an *explanation* is a description of the operation of a model, either in general or for a particular test vector, which covers the mechanism used to relate inputs to outputs (e.g. Harrison et al. (2017); Doshi-Velez and Kim (2017)). Explanations contrast in computer science to the inability to provide them and systems which cant be explained are sometimes viewed as lacking a deterministic connection between inputs and outputs.

However, in philosophy, explanations define the outline of what is and is not intelligible. Any action or entity which can be explained can also be ascribed meaning, while anything without an explanation cannot stem from an agent with intentions (Rosenberg (2015); Stone (2009)).

By contrast, the social sciences have a robust notion of how explanations should behave (Miller (2017)): explanations should be causal, explaining why an outcome was reached or an event occurred; they should be contrastive, explaining why event X happened over event Y; they should be selected, meaning they should be based on a few key causes rather than complete descriptions of a mechanism; and they should be social, meaning that they are meant to transfer knowledge about the system they are explaining.

Policymakers are elevating explanations to the status of individual rights in several jurisdictions (Selbst and Powles, 2017), but often without defining what would constitute an explanation or how it might be achieved. Here, lawmakers can learn what different communities require of explanations and what tools might exist to meet those requirements, so that practitioners can be given clarity on any new rules and enforcement or oversight bodies can act in accordance with the political intent driving the new-found focus on explanations.

6. Conclusion

This work situates how different research and practice communities approach and describe problems in the fairness, accountability, and transparency of software systems and presents a description of the lexicon conflated among different fields. By enabling those of different backgrounds to recognize when terms are being overloaded, this tutorial educates its audience to avoid speaking to members of other disciplines at cross purposes.

Acknowledgments

We are grateful to Deirdre Mulligan for her mentorship of this project and her deep knowledge of the literature.

References

Elizabeth S Anderson. What is the point of equality? *Ethics*, 109(2):287–337, 1999.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-\in-criminal-sentencing.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. *PMLR*, 81: 149–159, February 2018.

Robert Brauneis and Ellen P Goodman. Algorithmic transparency for the smart city. *Yale J. of Law and Tech.*, 2018. Forthcoming.

Julie Brill. Scalable approaches to transparency and accountability in decisionmaking algorithms, 2015. Remarks at the NYU Conference on Algorithms and Accountability.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Deven Desai and Joshua A. Kroll. Trust but verify: A guide to algorithms and the law. *Harvard J. of Law and Tech.*, 31, 2018.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity, 2016.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

Cynthia Dwork and Deirdre K Mulligan. It's not privacy, and it's not fair. *Stan. L. Rev. Online*, 66:35, 2013.

Ronald Dworkin. What is equality? *Philosophy & public affairs*, pages 185–246, 1981.

Andrew Gelman. Ethics and statistics: Honesty and transparency are not enough. *Chance*, 30 (1):37–39, 2017.

Anthony G Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. *Cal. L. Rev.*, 94:945, 2006.

Andreas Haeberlen, Petr Kouznetsov, and Peter Druschel. Peerreview: Practical accountability for distributed systems. In *ACM SIGOPS Operating Systems Review*, volume 41:6, pages 175–188. ACM, 2007.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Neural Information Processing Symposium*, 2016.

Brent Harrison, Upol Ehsan, and Mark O Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. *arXiv preprint arXiv:1702.07826*, 2017.

Anna Lauren Hoffman. Data violence: Dignity, discrimination, and algorithmic identity, 4 2017. URL http://sched.co/81F9. Keynote delivered at the Electronic Resources & Libraries Conference (ER&L), Austin, TX.

Chris Hoofnagle. How the fair credit reporting act regulates big data. *Future of Privacy Forum Workshop on Big Data and Privacy: Making Ends Meet*, 2013.

J.G. Jabbra and O.P. Dwivedi, editors. *Public Service Accountability: A Comparative Perspective*. Kumarian Press library of management for development. Kumarian Press, 1989. ISBN 9780931816413.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Joshua A. Kroll. *Accountable Algorithms*. PhD thesis, Princeton University, 2015.

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review (to appear)*, 165, 2017.

Ralf Küsters, Tomasz Truderung, and Andreas Vogt. Accountability: definition and relationship to verifiability. In *Proc. 17th ACM conf. Computer and Communications Security*, pages 526–535. ACM, 2010.

Amanda Levendowski. How copyright law can fix artificial intelligence's implicit bias problem. *Wash. L. Rev.*, 2018. Forthcoming.

Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's disparate impact require disparate treatment? *arXiv preprint arXiv:1711.07076*, 2017.

Edward Miguel, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M Esterling, Alan Gerber, Rachel Glennerster, Don P Green, Macartan Humphreys, Guido Imbens, et al. Promoting transparency in social science research. *Science*, 343(6166):30–31, 2014.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

Helen Nissenbaum. Accountability in a computerized society. *Science and engineering ethics*, 2(1):25–42, 1996.

Safiya Umoja Noble. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

Frank Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

Alexander Rosenberg. *Philosophy of social science*. Hachette UK, 2015.

Thomas Scanlon. When does equality matter? *unpublished paper*, page 15, 2004.

Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

Rajiv C Shah and Jay P Kesan. Manipulating the governance characteristics of code. *info*, 5 (4):3–9, 2003.

Daniel J Solove and Paul Schwartz. *Information privacy law*. Wolters Kluwer Law & Business, 2014.

Peter Stone. Rationality, intelligibility, and interpretation. *Rationality and Society*, 21(1): 35–58, 2009.

United States Department of Health, Education, and Welfare. *Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens: Report*. MIT Press, 1973.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Conference on the World Wide Web*, pages 1171–1180, 2017.