7

Principles to Govern Regulation of Digital and Machine Evidence

ANDREA ROTH

I Introduction

Criminal prosecutions now routinely involve technologically sophisticated tools for both investigation and proof of guilt, from complex software used to interpret DNA mixtures, to digital forensics, to algorithmic risk assessment tools used in pre-trial detention, sentencing, and parole determinations. As Emily Silverman, Jörg Arnold, and Sabine Gless's Chapter 8 explains, these tools offer not merely routine measurements, but also "evaluative data" akin to expert opinions.[1] These new tools, in critical respects, are a welcome addition to less sophisticated or more openly subjective forms of evidence that have led to wrongful convictions in the past, most notably eyewitness identifications, confessions, and statements of source attribution using "first generation"[2] forensic disciplines of dubious reliability, such as bite marks.[3]

Nonetheless, this new generation of evidence brings new costs and challenges. Algorithmic tools offer uniformity and consistency, but potentially at the expense of equitable safety valves to correct the unjust results that would otherwise flow from mechanistic application of rules. Such tools also may appear more reliable or equitable than they are, as fact-finders fail to identify sources of error or bias because the tools appear objective and are shrouded in black box secrecy. Even with greater transparency, some results, such as the decisions of deep neural networks engaged in deep

[1] See generally Chapter 8 in this volume.

[2] See Erin Murphy, "The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence" (2007) 95:3 *California Law Review* 721 ["New Forensics"] (comparing "first-generation" techniques, such as tool-marks and handwriting, to "second-generation" techniques, such as DNA and digital evidence).

[3] See generally Innocence Project, "DNA Exonerations in the United States (1989–2020)," https://innocenceproject.org/dna-exonerations-in-the-united-states/ (noting numerous exonerations in cases involving mistaken eyewitnesses, false confessions, and embellished forensic evidence).

learning, will not be fully explainable without sacrificing the very complexity that is the ostensible comparative advantage of artificial intelligence (AI). The lack of explainability as to the method and results of sophisticated algorithmic tools has implications for accuracy, but also for public trust in legal proceedings and participants' sense of being treated with dignity. As Sara Sun Beale and Haley Lawrence note in their Chapter 6 of this volume, humans have strong reactions to certain uses of robot "testimony" in legal proceedings.[4] Absent proper regulation, such tools may jeopardize key systemic criminal justice values, including the accuracy expressed by convicting the guilty and exonerating the innocent, fairness, public legitimacy, and softer values such as mercy and dignity.

In furtherance of these systemic goals, this chapter argues for four overarching principles to guide the use of digital and machine evidence in criminal justice systems: a right to front-end safeguards to minimize error and bias; a right of access both to government evidence and to exculpatory technologies; a right of contestation; and a right to an epistemically competent fact-finding process that keeps a human in the loop. The chapter offers legal and policy proposals to operationalize each principle.

Three caveats are in order. First, this chapter draws heavily on examples from the United States, a decentralized and adversarial system in which the parties themselves investigate the case, find witnesses, choose which evidence to introduce, and root out truth through contestation. Sabine Gless has described the many differences between the US and German approaches to machine evidence, distinguishing their adversarial and inquisitorial approaches, respectively.[5] Nonetheless, the principles discussed here are relevant to any system valuing accuracy, fairness, and public legitimacy. For example, although many European nations have a centralized, inquisitorial system, proposed EU legislation evinces concern over the rights of criminal defendants vis-à-vis AI systems, specifically the potential threat AI poses to a "fair" trial, the "rights of the defense," and the right to be "presumed innocent," as guaranteed by the EU Charter of Fundamental Rights.[6] As noted in Chapter 10 of

[4] See Chapter 6 in this volume.

[5] Sabine Gless, "AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials" (2020) 51:2 *Georgetown Journal of International Law* 195 ["AI in the Courtroom"].

[6] EU Charter of Fundamental Rights, 2000 (came into force in 2009), Title VI, Arts. 47–48; see also Artificial Intelligence Act, European Union (proposed April 21, 2021), COM(2021) 206 final 2021/0106, Explanatory Memorandum s. 3.5, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

this volume by Bart Custers and Lenneke Stevens, European nations are facing similar dilemmas when it comes to the regulation of digital evidence in criminal cases.[7]

The second caveat is that digital and machine evidence is a wide-ranging and definitionally vague concept. Erin Murphy's Chapter 9 in this volume offers a helpful taxonomy of such evidence that explains its various uses and characteristics, which in turn determine how such evidence implicates the principles in Section II.[8] Electronic communications and social media, e.g., implicate authentication and access concerns, but not so much the need for equitable safety valves in automated decision-making. Likewise, biometric identifiers may raise more privacy concerns than use of social media posts as evidence. The key characteristics of digital evidence as cataloged by Murphy also affect which principles are implicated. For example, data created by a private person, and possessed by Facebook, might implicate the right to exculpatory information and the Stored Communications Act,[9] while resiliency or lack of data such as body-worn camera footage might require the state to adopt more stringent preservation and storage measures, and to allow defendants access to e-discovery tools. So long as the principles are followed when they apply, the delivery of justice can be enhanced rather than jeopardized by digital and machine proof.

The third caveat is that this chapter does not write on a blank slate in setting forth principles to govern the use of technology in rendering justice. A host of disciplines and governing bodies have adopted principles for "ethical or responsible" use of AI, from the US Department of Defense to the Alan Turing Institute to the Council of Europe. Recent meta-studies of these various sets of principles have identified recurring values, such as beneficence, autonomy, justice, explainability, transparency, fairness, responsibility, privacy, expert oversight, stakeholder-driven legitimacy, and "values-driven determinism."[10] More specifically, many countries

[7] See Chapter 10 in this volume (exploring the shift toward digital evidence in Dutch criminal courts).

[8] See Chapter 9 in this volume. Erin Murphy divides "technological evidence" into location trackers, electronic communications and social media, historical search or cloud or vendor records, "Internet of Things" and smart tools, surveillance cameras and visual imagery, biometric identifiers, and analytical software tools.

[9] 18 United States Code [18 USC], §§2701–2712.

[10] See e.g. Luciano Floridi & Josh Cowls, "A Unified Framework of Five Principles for AI in Society" (2019) 1:1 *Harvard Data Science Review* (examining forty-seven principles promulgated since 2016, which map onto beneficence, non-maleficence, autonomy, justice,

already have a detailed legal framework to govern criminal procedure. In the United States, e.g., criminal defendants already have a constitutional right to compulsory process, to present a defense, to be confronted with the witnesses against them, to a verdict by a human jury, and to access to experts where necessary to a defense. But these rights were established at a time when cases largely depended on human witnesses rather than machines. The challenge here is not so much to convince nations in the abstract to allow a right to contest automated decision-making, but to explain how existing rights, such as the right of confrontation or right to pre-trial disclosure of the bases of expert testimony, might apply to this new type of evidence.

II The Principles

Principle I: The digital and machine evidence used as proof in criminal proceedings should be subject to front-end development and testing safeguards designed to minimize error and bias.

Principle I(a): Jurisdictions should acknowledge the heightened need for front-end safeguards with respect to digital and machine evidence, which cannot easily be scrutinized through case-specific, in-trial procedures.

To understand why the use of digital and machine evidence merits special front-end development and testing safeguards that do not apply to all types of evidence, jurisdictions should acknowledge that the current real-time trial safeguards built for human witnesses, such as cross-examination, are not as helpful for machine-generated proof.

and explicability); Anna Jobin, Marcello Ienca, & Effy Vayena, "The Global Landscape of AI Ethics Guidelines" (2019) 1:9 *Nature Machine Intelligence* 389–399 (reviewing 84 documents, which centered around transparency, justice and fairness, non-maleficence, responsibility, and privacy); Daniel Greene, Anna Lauren Hoffmann, & Luke Stark, "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning" (paper delivered at the Proceedings of the 52nd Hawaii International Conference on System Sciences, January 8, 2019), cited in Samuele Lo Piano, "Ethical Principles in Machine Learning and Artificial Intelligence: Cases from the Field and Possible Ways Forward" (2020) 7:1 *Humanities and Social Sciences Communications*, Article 9 (collecting meta-studies). The Council of Europe's 2020 Resolution on AI also includes these values, specifically mentioning "transparency, including accessibility and explicability," "justice and fairness," and "human responsibility for decisions." See Council of Europe, "Council of Europe and Artificial Intelligence," www.coe.int/en/web/artificial-intelligence.

A critical goal of any criminal trial is to ensure verdict accuracy by minimizing the chance of the fact-finder drawing the wrong inferences from the evidence presented. There are several different levers a system could use to combat inferential error by a jury. First, the system could exclude unreliable evidence so that the jury never hears it. Second, the system could implement front-end design and production safeguards to ensure that evidence is as reliable as it can be when admitted, or that critical contextual information about its probative value is developed and disclosed when the fact-finder hears it. Third, the system could allow parties themselves to explore and impeach, or attack the credibility/reliability of the evidence. Fourth, the system could adopt proof standards that limit the fact-finder's ability to render a verdict absent a proof threshold such as beyond a reasonable doubt, or type or quantum of evidence.

For better or worse, the American system of evidence pursues accuracy almost entirely through trial and back-end safeguards, the third and fourth levers described above. Although the United States still clings to the rule excluding hearsay, understood as out-of-court statements offered for their truth, that rule has numerous exceptions. And while US jurisdictions used to have stringent competence requirements for witnesses, these have given way to the ability to impeach witnesses once they testify or once their hearsay is admitted.[11] The parties conduct such impeachment through cross-examination, physical confrontation, and admission of extrinsic evidence such as a witness's prior convictions or inconsistent statements. In addition, the United States has back-end proof standards to correct for unreliable testimony, such as corroboration requirements for accomplice testimony and confessions. The US system has a similarly lenient admission standard with regard to physical evidence, requiring only minimal proof that an item such as a document or object is what the proponent says it is.[12]

Nonetheless, there are particular types of witness testimony that do require more front-end safeguards, ones that could work well for digital and machine evidence too. One example is eyewitness identifications. If an identification is conducted under unnecessarily suggestive circumstances, a US trial court, as a matter of constitutional due process, must conduct

[11] See e.g. Federal Rules of Evidence, United States (as amended on December 1, 2020) [Federal Rules of Evidence], Rules 602 (liberal competence standard), 806 (allowing impeachment of hearsay declarants), 608–609 (allowing impeachment by character-for-dishonesty evidence), and 613 and 801(d) (impeachment by inconsistent statements).

[12] See e.g. Federal Rules of Evidence, note 11 above, Rules 901 and 902 (imposing minimal authentication requirements).

a hearing to determine whether the identification is sufficiently reliable to be admitted against the defendant at trial.[13] Moreover, some lower US courts subject identification testimony to limits or cautionary instructions at trial, unless certain procedures were used during the identification, to minimize the risk of suggestivity.[14] Likewise, expert testimony is subjected to enhanced reliability requirements that question whether the method has been tested, has a known error rate, has governing protocols, and has been subject to peer review.[15] To a lesser extent, *confession* evidence is also subject to more stringent front-end safeguards, such as the requirement in some jurisdictions that stationhouse confessions be videotaped.[16]

The focus on front-end safeguards in these specific realms is not a coincidence. Rather, it stems from the fact that the problems with such testimony are largely cognitive, subconscious, or recurring, rather than a matter of one-off insincerity, and therefore not meaningfully scrutinized solely through cross-examination and other real-time impeachment methods.[17] These categories of testimony bear some of the same process-like characteristics that make digital and machine evidence difficult to scrutinize through cross-examination alone.

Even more so than these particular types of human testimony, digital and machine evidence bear characteristics that call for robust front-end development and testing safeguards before it gets to the courtroom. First, the programming of the algorithms that drive the outputs of many of the categories of proof discussed by Erin Murphy, including location trackers, smart tools, and analytical software tools, does not necessarily change from case to case.[18] Repeatedly-used software can be subject to testing to

[13] *Manson* v. *Brathwaite*, 432 U.S. 98 (1977).

[14] See e.g. *State* v. *Henderson*, 27 A.3d 872, 878 (NJ 2011) (establishing protocols for eyewitness identification procedures).

[15] See *Daubert* v. *Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993) (setting forth a non-exhaustive list of factors trial courts should use in determining the scientific validity of an expert method). A minority of US state jurisdictions continue to adhere to the alternative *Frye* test, that looks to whether novel scientific methods are "general[ly] accept[ed]" in the scientific community. See *Frye* v. *United States*, 293 F. 1013 (DC Cir. 1923).

[16] See e.g. G. Daniel Lassiter, Andrew L. Geers, Ian M. Handley *et al.*, "Videotaped Interrogations and Confessions: A Simple Change in Camera Perspective Alters Verdicts in Simulated Trials" (2002) 87:5 *Journal of Applied Psychology* 867 at 867.

[17] See Edward Cheng & Alexander Nunn, "Beyond the Witness: Bringing a Process Perspective to Modern Evidence Law" (2019) 97:6 *Texas Law Review* 1077 ["Beyond the Witness"]; see also Jules Epstein, "The Great Engine that Couldn't: Science, Mistaken Identifications, and the Limits of Cross-Examination" (2007) 36:3 *Stetson Law Review* 727.

[18] See e.g. "New Forensics", note 2 above (noting this aspect of "second-generation" forensic techniques like DNA).

determine its accuracy under various conditions. Second, unlike eyewitnesses and confessions, where the declarant in some cases might offer significant further context through testimony, little further context can be gleaned from in-court scrutiny of any of the categories of proof Murphy describes.[19] To be sure, a programmer or inputter could take the stand and explain some aspects of a machine's output in broad strokes. But the case-specific "raw data," "measurement data," or "evaluative data"[20] of the machine is ultimately the product of the operation of the machine and its algorithms, not the programmer's own mental processes, and it is the machine's and algorithm's operation that must also be scrutinized. In short, the accoutrements of courtroom adversarialism, such as live cross-examination, are hardly the "greatest legal engine ever invented for the discovery of truth"[21] of the conveyances of machines.

Principle I(b): Jurisdictions should implement and enforce, through admissibility requirements, certain minimal development and testing procedures for digital and machine evidence.

Several development and testing safeguards should be implemented for any software-driven system whose results are introduced in criminal proceedings. The first is robust, independent stress testing of the software. Such standards are available,[22] but are typically not applied, at least in the United States, to software created for litigation. For example, a software expert reviewing the code of the Alcotest 7110, a breath-alcohol machine used in several US states, found that it would not pass industry standards. He documented 19,500 errors, nine of which he believed "could ultimately [a]ffect the breath alcohol reading."[23] A reviewing court held that such errors did not merit excluding the

[19] See ibid.; see also Chapter 9 in this volume.

[20] See Chapter 8 in this volume. The chapter defines "raw data" as data produced by a machine without any processing, "measurement data" as data produced by a machine after rudimentary calculations, and "evaluative data" as data produced by a machine according to sophisticated algorithmic methods that cannot be reproduced manually.

[21] Prominent American evidence scholar John Henry Wigmore famously described cross-examination in this way, see John Henry Wigmore, *Evidence in Trials at Common Law*, vol. 5 (Boston, MA: Little, Brown & Co., 1974) at 32, s. 1367.

[22] See e.g. Declaration of Nathaniel Adams, *People* v. *Hillary*, No. 2015–15 (New York County Court of St Lawrence, May 27, 2016) at 1–2 (on file with author) (listing citations to several governing bodies that have come together to promulgate industry standards for software development and testing).

[23] See Supplemental Findings and Conclusions of Remand Court at 11, *State* v. *Chun*, No. 58,879 (NJ November 14, 2007), www.nj-dmv-dwi.com/state-v-chun-alcotest-litigation/.

reading, in part because the expert could not say with "reasonable certainty" that the errors caused a false reading in the case at hand,[24] but the court did require modifications of the program for future use.[25] In addition, Nathaniel Adams, a computer scientist and expert in numerous criminal cases in the United States, has advocated for forensic algorithms to be subject to the industry-standard testing standards of the Institute of Electrical and Electronic Engineers (IEEE).[26] Adams notes that STRMix, one of the two primary probabilistic genotyping programs used in the United States, had not been tested by a financially independent entity,[27] and the program's creators have disclosed more than one episode of miscodes potentially affecting match statistics, thus far, in ways that would underestimate but not overestimate a match probability.[28] Professor Adams' work helped to inspire a recent bill in the US Congress, the Justice in Forensic Algorithms Act of 2021, which would subject machine-generated proof in criminal cases to more rigorous testing, along with pre-trial disclosure requirements, defense access, and the removal of trade secret privilege from proprietary code.[29] And exclusion aside, a rigorous software testing requirement reduces the chance of misleading or false machine conveyances presented at trial.

Jurisdictions should also enact mandatory testing and operation protocols for machine tools used to generate evidence of guilt or innocence, along the lines currently used for blood-alcohol breath-testing

[24] Ibid.

[25] See *State* v. *Chun*, 943 A.2d 114, 129–30 (NJ 2008); see also Robert Garcia, "'Garbage in, Gospel Out': Criminal Discovery, Computer Reliability, and the Constitution" (1991) 38:5 *UCLA Law Review* 1043 at 1088 (citing GAO report finding deficiencies in software used by Customs Office to record license plates, and investigations of failures of IRS's computer system).

[26] See e.g. Nathaniel Adams, "What Does Software Engineering Have to Do with DNA?" (2018) May Issue *NACDL The Champion* 58 ["Software Engineering"] (arguing that software should be subject to industry-standard IEEE-approved independent software testing); Andrea Roth, "Machine Testimony" (2017) 126:7 *Yale Law Journal* 1972 ["Machine Testimony"] at 2023 (arguing for independent software testing as admissibility requirement).

[27] "Software Engineering", note 26 above.

[28] See *Final Report – Variation in STRMix Regarding Calculation of Expected Heights of Dropped Out Peaks* (STRMix, July 4, 2016) at 1–2 (on file with author) (acknowledging coding errors, but noting that errors would only underestimate the likelihood of contribution). Of course, an error underestimating the likelihood of contribution might also be detrimental to a factually innocent defendant in certain cases, such as where the defense alleges a third-party perpetrator.

[29] See United States, Bill HR 2438, Justice in Forensic Algorithms Act of 2021, 117th Cong., 2021, www.govtrack.us/congress/bills/117/hr2438.

equipment.[30] Such requirements need not be a condition of admission; in the breath-alcohol context, the failure to adhere to protocols goes to weight, not admissibility.[31] Even so, the lack of validation studies showing an algorithm's accuracy under circumstances relevant to the case at hand should, in some cases, be a barrier to admissibility. Jurisdictions should subject the conclusions of machine experts to validity requirements at the admissibility stage, similar to those imposed on experts at trial. Currently, the *Daubert* and *Frye* reliability/general acceptance requirements apply only to human experts; if the prosecution introduces machine-generated proof without a human interlocutor, the proof is subject only to general authentication and relevance requirements.[32]

Requiring the proponent to show that the algorithm is fit for purpose through developmental and internal validation before offering its results is key not merely for algorithms created for law enforcement but for algorithms created for commercial purposes as well. For example, while Google Earth results have been admitted as evidence of guilt with no legal scrutiny of their reliability,[33] scientists have conducted studies to determine its error rate with regard to various uses.[34] While error is inevitable in any human or machine method, this type of study should be a condition of admitting algorithmic proof.[35]

Such testing need not necessarily require public disclosure of source code or other levels of transparency that could jeopardize intellectual property interests. Instead, testing algorithms for forensic use could be done in a manner similar to testing of potentially patentable pharmaceuticals by

[30] See e.g. Conforming Products List of Evidential Breath Alcohol Measurement Devices, 2012, 77 Fed. Reg. 35,747, 35,748 (prohibiting states from using machines except those approved by the National Highway Transportation Safety Administration).

[31] See e.g. *People* v. *Adams*, 131 Cal. Rptr. 190, 195 (Ct. App. 1976) (holding that a failure to calibrate breath-alcohol equipment went only to weight).

[32] See e.g. *People* v. *Lopez*, 286 P.3d 469, 494 (Cal. 2012) (admitting results of gas chromatograph, without testimony of expert); "Machine Testimony", note 26 above, at 1989–1990 (explaining that the hearsay rule does not apply to machines, heightening the need for alternative forms of scrutiny).

[33] See e.g. *United States* v. *Lizarraga-Tirado*, 789 F.3d 1107, 1109 (9th Cir. 2015) (admitting Google Earth "pin" associated with GPS coordinates as evidence that defendant had been arrested on the US side of the US–Mexico border for purposes of an illegal re-entry prosecution).

[34] See e.g. Shawn Harrington, Joseph Teitelman, Erica Rummel *et al.*, "Validating Google Earth Pro as a Scientific Utility for Use in Accident Reconstruction" (2017) 5:2 *SAE International Journal of Transport Safety* 135.

[35] Cf. "Beyond the Witness", note 17 above (arguing that process-based evidence should be subject to testing to determine error rate).

the US Food and Drug Administration.[36] Others have made the point that scrutiny by "entrusted intermediate parties," behind closed doors, would avoid any financial harm to developers.[37] Of course, for algorithms that are open source, such concerns would be lessened.

One limit on validation studies as a guarantor of algorithmic accuracy is that most studies do not speak to whether an algorithm's reported score or statistic, along a range, is accurate. Studies might show that a software program boasts a low false positive rate in terms of falsely labeling a non-contributor as a contributor to a DNA mixture, but not whether its reported likelihood ratio might be off by a factor of ten. As two DNA statistics experts explain, there is no "ground truth" against which to measure such statistics:

> Laboratory procedures to measure a physical quantity such as a concentration can be validated by showing that the measured concentration consistently lies with an acceptable range of error relative to the true concentration. Such validation is infeasible for software aimed at computing a [likelihood ratio] because it has no underlying true value (no equivalent to a true concentration exists). The [likelihood ratio] expresses our uncertainty about an unknown event and depends on modeling assumptions that cannot be precisely verified in the context of noisy [crime scene profile] data.[38]

But systems are not helpless in testing the accuracy of algorithm-generated credit scores or match statistics. Rather, such results must be scrutinized using other methodologies, such as more complex studies that go beyond simply determining false positive rates, stress testing of software, examination of source code by independent experts, and assessment of whether various inputs, such as assumptions about the values of key variables, are appropriate.

Principle I(c): Jurisdictions should explicitly define what is meant by algorithmic fairness for purposes of testing for, and guarding against, bias.

Algorithms should also be tested for bias. The importance of avoiding racial and other bias in algorithmic decision-making is perhaps obvious,

[36] See e.g. Andrew Tutt, "An FDA for Algorithms" (2017) 69:1 *Administrative Law Review* 83 (suggesting that such a body could prevent problematic algorithms from going to market).

[37] Paul B. de Laat, "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" (2018) 31:4 *Philosophy & Technology* 525.

[38] Christopher D. Steele & David J. Balding, "Statistical Evaluation of Forensic DNA Profile Evidence" (2014) 1:1 *Annual Review of Statistics and Its Application* 361 at 380.

given that fairness is an explicitly stated value in nearly all promulgated AI standards in the meta-studies referenced in the introduction to this chapter. In addition, racial, gender, and other kinds of bias might trigger legal violations as well as ethical or policy concerns. To be sure, the Equal Protection Clause of the Fourteenth Amendment to the US Constitution guards only against state action that intentionally treats people differently because of a protected status, but if an algorithm simply has a disparate impact on a group, it will likely not be viewed as an equal protection violation. However, biased algorithms used in jury selection could violate the requirement that petit juries be drawn from a fair cross section of the population, and biased algorithms used to prove dangerousness or guilt at trial could violate statutory anti-discrimination laws or reliability-based admissibility standards.

In one highly publicized example of algorithmic bias from the United States, Pro Publica studied Northpointe's post-trial risk assessment tool Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) and determined that the false positive rates, i.e., rates of those labeled "dangerous," but who did not reoffend, for Black subjects was much higher than for White subjects.[39] At the same time, however, other studies, including by Northpointe itself, noted that the algorithm is, in fact, racially non-biased if the metric is whether race has any predictive value in the model in determining dangerousness.[40] As Northpointe notes, Black and White subjects with the same risk score present the same risk of reoffending under the model.[41] The upshot was not that Pro Publica was wrong in noting the differences in false positive rates; it was that Pro Publica judged the algorithm's racial bias by only one particular measure.

The COMPAS example highlights the problems of testing algorithms for fairness without defining terms. As others have explained, it is impossible to have both equal false positive rates and predictive

[39] See Jeff Larson, Surya Mattu, Lauren Kirchner *et al.*, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica* (May 23, 2016), www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[40] See "Response to ProPublica: Demonstrating Accuracy, Equity, and Predictive Parity," *Northpointe Research Department* (July 8, 2016), www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/ ["Response to ProPublica"]; Jon Kleinberg, Sendhil Mullainathan, & Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *Cornell University* (November 17, 2016), arxiv.org/abs/1609.05807v2 (arguing that algorithms like COMPAS cannot simultaneously satisfy all three possible means of measuring algorithmic fairness, and that it has predictive parity even with different false positive rates).

[41] "Response to ProPublica", note 40 above.

parity where two groups have different base rates.[42] So, in determining whether the algorithm is biased, one needs to decide which measure is the more salient indicator of the type of bias the system should care about. Several commentators have noted possible differences in definitions of algorithmic fairness as well.[43] Deborah Hellman argues that predictive parity alone is an ill-suited measure of algorithmic fairness because it relates only to beliefs, not outcomes.[44] In Hellman's view, a disparate false positive rate between groups is highly relevant to proving, though not dispositive of, normatively troubling unfairness.[45] While not all jurisdictions will agree with Hellman's take, the point is that algorithm designers should be aware of different conceptions of fairness, be deliberate in choosing a metric, and ensure that algorithms in criminal proceedings are fair under that metric. Jurisdictions could require what Osagie Obasogie has termed "racial impact statements" in the administrative law context,[46] to determine the effect of a shift in decision-making on racial groups. The Council of Europe has made a similar recommendation, calling on states to conduct "human rights impact assessments of AI applications" to assess "risks of bias/discrimination … with particular attention to the situation of minorities and vulnerable and disadvantaged groups."[47]

Finally, in determining algorithmic fairness, decision-makers should judge algorithms not in a vacuum, but against existing human-driven decision-making processes. For example, court reporters have been known to mistakenly transcribe certain dialects, such as African American Vernacular English (AAVE), in ways that matter to fact-finding

[42] See e.g. Richard Berk, Hoda Heidari, Shahin Jabbari *et al*., "Fairness in Criminal Justice Risk Assessments: The State of the Art" (2021) 50:1 *Sociological Methods & Research* 3 (explaining that these two types of fairness are incompatible).

[43] See e.g. Dana Pessach & Erez Schmueli, "Algorithmic Fairness," *Cornell University* (January 21, 2020), https://arxiv.org/abs/2001.09784 (noting that COMPAS offered certain types of predictive parity, but that the odds of being predicted dangerous were worse for African-Americans than White subjects).

[44] Deborah Hellman, "Measuring Algorithmic Fairness" (2020) 106:4 *Virginia Law Review* 811.

[45] Ibid. at 840–841.

[46] Osagie K. Obasogie, "The Return of Biological Race? Regulating Race and Genetics Through Administrative Agency Race Impact Assessments" (2012) 22:1 *Southern California Interdisciplinary Law Journal* 1.

[47] "Justice by Algorithm – The Role of Artificial Intelligence in Policing and Criminal Justice Systems," Doc. 15156, report of the Committee on Legal Affairs and Human Rights, Resolution 2342 (Council of Europe, Parliamentary Assembly, 2020), https://pace.coe.int/en/files/28805/html ["Justice by Algorithm"].

in criminal proceedings.[48] If an AI system were to offer a lower, even if non-zero, error rate with regard to mistranscriptions of AAVE, the shift toward such systems, at least a temporary one subject to continued testing and oversight, might reduce, rather than exacerbate, bias.[49]

Principle II: Before trial or other relevant proceeding, the parties should have meaningful and equitable access to digital and machine evidence material to the proceeding, including exculpatory technologies and data.

Principle II(a): Pretrial disclosure requirements related to expert testimony should apply to digital and machine conveyances that, if asserted by a human, would be subject to such requirements.

Because digital and machine evidence cannot be cross-examined, parties cannot use the in-court trial process itself to discover the infirmities of algorithms or possible flaws in their results or opinions. As Edward Cheng and Alex Nunn have noted, enhanced pre-trial discovery must in part take the place of in-court discovery with regard to process-based evidence like machine conveyances.[50] Such enhanced discovery already exists in the United States for human experts, precisely because in-court examination alone is not a meaningful way for parties to understand and prepare to rebut expert testimony. Specifically, parties in criminal cases are entitled by statute to certain information with regard to expert witnesses, including notice of the basis and content of the expert's testimony and the expert's qualifications.[51] Disclosure requirements in civil trials are even more onerous, requiring experts to prepare written reports that include the facts or data relied on.[52] Moreover, proponents of expert testimony must not discourage experts from speaking with the opposing party,[53] and in criminal trials, proponents must also disclose certain prior statements, or Jencks material, of witnesses after they testify.[54] These requirements

[48] See e.g. Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock *et al.*, "Testifying While Black: An Experimental Study of Court Reporter Accuracy in Transcription of African American English" (2019) 95:2 *Language: Linguistic Society of America* 216.

[49] Whether AI voice-recognition-driven court reporting systems are more accurate than human stenographers remains to be seen.

[50] "Beyond the Witness", note 17 above.

[51] See e.g. Federal Rules of Criminal Procedure, United States (as amended December 1, 2022) [Federal Rules of Criminal Procedure], Rule 16(a)(1)(G).

[52] See Federal Rules of Criminal Procedure, note 51 above, Rule 26(a)(2)(B)(ii).

[53] See e.g. *Gregory* v. *United States*, 369 F.2d 185, 188 (DC Cir. 1966) ("Both sides have an equal right, and should have an equal opportunity, to interview [state witnesses]").

[54] See e.g. 18 USC, note 9 above, Jencks Act, 18 USC §3500(b).

also facilitate parties' ability to consult with their own experts to review the opposing party's evidence or proffered expert testimony.

Using existing rules for human experts as a guide, jurisdictions should require that parties be given access to the following:

(1) The evidence and algorithms themselves, sufficient to allow meaningful testing of their assumptions and running the program with different inputs. One probabilistic genotyping software company, TrueAllele, offers defendants access to its program, with certain restrictions, albeit only for a limited time and without the source code.[55] This sort of "black box tinkering" not only allows users to "confront" the code "with different scenarios," thus "reveal[ing] the blueprints of its decision-making process,"[56] but also approximates the posing of a hypothetical to a human expert. Indeed, the ability to tinker might be just as important as access to source code; data science scholars have written about the limits of transparency and the superior promise of reverse engineering in understanding how inputs relate to outputs.[57] Along these lines, Jennifer Mnookin has argued that a condition for admissibility of computer simulations should be that "their key evidence-based inputs are modifiable," allowing the opposing party to "test the robustness of the simulation by altering the factual assumptions on which it was built and seeing how changing these inputs affects the outputs."[58]
(2) The training or software necessary to use or test the program. In the United States, criminal defendants have reported that certain trainings are off limits to non-law-enforcement; e.g., for using the Cellebrite program to extract digital evidence from a cell phone, or for using DNA genotyping software. Moreover, only certain defense experts are able to buy the software for their own use, and some academic researchers have been effectively denied research licenses to study proprietary forensic software. Instead, the defense and academic communities should presumptively be given a license to access

[55] See State's Response to Defense Motion to Compel, *State* v. *Fair*, No. 10-1-09274-5 (Wash. Sup. Ct. April 1, 2016) at 21 (representations made by TrueAllele as to defense access to its program).

[56] Maayan Perel & Niva Elkin-Koren, "Black Box Tinkering: Beyond Transparency in Algorithmic Enforcement" (2017) 69:5 *Florida Law Review* 181.

[57] Nick Diakopoulos, "Algorithmic Accountability Reporting: On the Investigation of Black Boxes" (2013) *Tow Center for Digital Journalism* 30, https://academiccommons.columbia.edu/doi/10.7916/D8ZK5TW2.

[58] Jennifer Mnookin, "Repeat Play Evidence: Jack Weinstein, 'Pedagogical Devices,' Technology, and Evidence" (2015) 64:2 *DePaul Law Review* 571 at 573.

to all software used by the government in generating evidence of guilt, to facilitate independent validity testing.

(3) A meaningful account of the assumptions underlying the machine's results or opinion, as well as the source code and prior output of software, where necessary to a meaningful understanding of those assumptions. Human experts can be extensively questioned both before and during trial, offering a way for parties to understand and refute their methods and conclusions. Digital and machine evidence cannot be questioned in the same way, but proponents should be required to disclose the same type of information about methods and conclusions that a machine expert witness would offer, if it could talk. Likewise, Article 15(1)(h) of General Data Protection Regulation (GDPR)[59] gives a data subject the right to know of any automated decision-making to which he is subject, and if so, the right to "meaningful information about the logic involved." While the GDPR may apply only to private parties rather than criminal prosecutions, the subject's dignitary interest in understanding the machine's logic would presumably be even greater in the criminal realm.

In particular, where disclosure of source code is necessary to meaningful scrutiny of the accuracy of machine results,[60] the proponent must allow access. As discussed in Principle I, source code might be important in particular to scrutinize scores or match statistics, where existing studies reveal only false positive rates. A jurisdiction should also require disclosure of prior output of the machine, covering the same subject matter as the machine results being admitted.[61] For human witnesses, such prior statements must be disclosed in many US jurisdictions to facilitate scrutiny of witness claims and impeachment by inconsistency. For machines, parties should have to disclose, e.g., the results of all prior runs of DNA

[59] General Data Protection Regulation, EU 2016, Regulation (EU) 2016/679 (with effect from May 25, 2018).

[60] See e.g. Andrew Morin, Jennifer Urban, Paul D. Adams *et al*., "Shining Light into Black Boxes" (2012) 336:6078 *Science* 159 at 159 ["Shining Light"] ("Common implementation errors in programs … can be difficult to detect without access to source code"); Erin E. Kenneally, "Gatekeeping Out of the Box: Open Source Software as a Mechanism to Assess Reliability for Digital Evidence" (2001) 6:13 *Virginia Journal of Law and Technology* 13 (arguing that access to source code is necessary to prevent or unearth many structural programming errors).

[61] See e.g. *United States* v. *Liebert*, 519 F.2d 542, 543, 550–51 (3d Cir. 1975) (entertaining the possibility that the defense was entitled to view the IRS program's prior reports of non-filers to determine their accuracy, but determining that access was not necessary to impeach the program).

software on a sample, all potentially matching reference fingerprints reported by a database using a latent print from a crime scene,[62] or calibration data from breath-alcohol machines.[63]

(4) Access to training data. Defendants and their experts should have access to underlying data used by the machine or algorithm in producing its results. In countries with inquisitorial as compared to adversarial systems, defendants should have access to "any data that is at the disposal of the court-appointed expert."[64] For example, for a machine-learning model labeling a defendant a "sexual psychopath" for purposes of a civil detention statute, the defendant should have access to the training dataset. Issues of privacy, i.e., the privacy of those in the dataset, have arisen, but are not insurmountable.[65]

To be sure, access alone does not guarantee that defendants will understand what they are given. But access is a necessary condition to allowing defendants to consult with experts who can meaningfully study the algorithms' performance and limits.

Principle II(b): Jurisdictions should not allow claims of trade secret privilege or statutory privacy interests to interfere with a criminal defendant's meaningful access to digital and machine evidence, including exculpatory technologies and data.

While creators of proprietary algorithms routinely argue that source code is a trade secret,[66] this argument should not shield code from discovery in a criminal case, where the code is material to the proceedings.[67] Of course, if proprietors can claim substantive intellectual property rights in their algorithms, those rights are still enforceable through licensing fees and civil lawsuits.

[62] State officials generally refuse defense requests for access to the other reported near matches, notwithstanding arguments that these matches might prove exculpatory. See generally Simon A. Cole, "More than Zero: Accounting for Error in Latent Fingerprint Identification" (2005) 95:3 *Journal of Criminal Law and Criminology* 985.

[63] Kathleen E. Watson, "COBRA Data and the Right to Confront Technology against You" (2015) 42:2 *North Kentucky Law Review* 375 at 381–382. But see *Turcotte* v. *Dir. of Revenue*, 829 S.W.2d 494, 496 (Mo. Ct. App. 1992) (holding that the state's failure to file timely maintenance reports on a breath-alcohol machine did not "impeach the machine's accuracy").

[64] "AI in the Courtroom", note 5 above, at 248.

[65] See e.g. Emiliano De Cristofaro, "An Overview of Privacy in Machine Learning," *Cornell University* (May 18, 2020), https://arxiv.org/abs/2005.08679.

[66] See generally Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System" (2018) 70:5 *Stanford Law Review* 1343.

[67] Ibid. (arguing that trade secrets doctrine should not apply in criminal cases).

Likewise, criminal defendants should have meaningful access to exculpatory digital and machine evidence, including the ability to subpoena witnesses who can produce such evidence in criminal proceedings where such evidence is material. Rebecca Wexler has explored the asymmetries inherent in US statutes such as the Stored Communications Act, which shields electronically stored communications from disclosure and has an exception for "law enforcement," but not for criminal defendants, however material the communications might be to establishing innocence. Such asymmetries are inconsistent not only with basic adversarial fairness, but arguably also with the Sixth Amendment compulsory process.[68]

Principle II(c): Jurisdictions should apply a presumption in favor of open-source technologies in criminal justice.

In the United States, the public has a constitutional right of access to criminal proceedings.[69] With regard to human witnesses, the public can hear the witnesses testify and determine the strength and legitimacy of the state's case. The public should likewise be recognized as a stakeholder in the development and use of digital and machine evidence in criminal proceedings. The Council of Europe's guidelines for use of AI in criminal justice embrace this concept, requiring Member States to "meaningfully consult the public, including civil society organizations and community representatives, before introducing AI applications."[70]

The most direct way to ensure public scrutiny of such evidence would be through open-source software. Scholars have discussed the benefits of open-source software in terms of facilitating "crowdsourcing"[71] and "ruthless public scrutiny"[72] as means of testing models and algorithms for hidden biases and errors. Others have gone further, arguing that software should

[68] See generally Rebecca Wexler, "Privacy Asymmetries: Access to Data in Criminal Defense Investigations" (2021) 68:1 *UCLA Law Review* 212.

[69] See *In re. Oliver*, 333 U.S. 257 (1948); Sixth Amendment to the US Constitution (right to a "public trial").

[70] "Justice by Algorithm", note 47 above, at 9.3.

[71] Cathy O'Neil, *Weapons of Math Destruction* (New York, NY: Crown Books, 2016) [*Weapons of Math Destruction*] at 211 (calling for "crowdsourcing campaigns" to offer feedback on errors and biases in datasets and models); see also Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (Cambridge, MA: Harvard University Press, 2015) at 208 (arguing for open source software in determining credit scores).

[72] Holly Doremus, "Listing Decisions under the Endangered Species Act: Why Better Science Isn't Always Better Policy" (1997) 75:3 *Washington University Law Quarterly* 1029 at 1138.

be open source whenever used in public law.[73] Public models would have the benefit of being "transparent" and "continuously updated, with both the assumptions and the conclusions clear for all to see."[74] States could encourage adoption of open-source software through drastic means, excluding output from adjudication, or more modest means, such as offering monetary incentives or prizes for development of open source replacements.

Principle II(d): Jurisdictions should make investigative technologies equally available to criminal defendants for potential exculpatory purposes, regardless of whether the state used the technology in a given case.

As Erin Murphy notes in Chapter 9 of this volume, defendants have two compelling needs with regard to digital and machine evidence: a meaningful chance to attack the government's proof, and a meaningful chance to discover and present "supportive defense evidence."[75] Just as both defendants and prosecutors have the ability to interview and subpoena witnesses, defendants should have an equal ability to wield new technologies that are paid for by the state when prosecutors seek to use them. If a defendant is accused of a crime based on what he believes to be a human analyst's erroneous interpretation of a complex DNA mixture, the defendant should be given the ability to use a probabilistic genotyping program, like TrueAllele, to attack these results. Of course, this access would be costly, and might reasonably be denied in cases where it bears no relevance to the defense, as determined *ex parte* by a judge. But if defendants have a due process right to access to defense experts where critical to their defense,[76] they should have such a right of access to exculpatory algorithms as well.

Principle III: Criminal defendants should have a meaningful right of contestation with respect to digital and machine evidence including, at a minimum, a right to be heard on development and testing procedures and meaningful access to experts.

Much has been written about a right of contestation by data subjects with regard to results of automated decision-making processes.[77] In the

[73] "Shining Light", note 60 above (arguing for open-source software for public law uses).
[74] *Weapons of Math Destruction*, note 71 above.
[75] See Chapter 9 in this volume.
[76] See *Ake* v. *Oklahoma*, 470 U.S. 68 (1986).
[77] See e.g. *Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems* (Council of Europe, Committee of Ministers, 2020) at 9, 13 ("[a]ffected individuals and groups should be afforded effective means to contest relevant determinations and decisions … [which] should include an opportunity to be heard, a thorough review of the decision and the possibility to obtain a non-automated decision");

US criminal context, defendants already enjoy, at least in theory, a right to present a defense, encompassing a cluster of rights, including the right to be confronted by the witnesses against them, to testify in their own defense, and to subpoena and present witnesses in their favor. In the United States, a criminal defendant's right of contestation essentially encompasses everything already discussed with regard to access to the state's evidence, as well as to some exculpatory electronic communications. In addition, the US Supreme Court has held that the right to present a defense exists even where the government presents scientific evidence of guilt that a trial judge might deem definitive. The fact that an algorithm offers compelling evidence of guilt cannot preclude a defendant from offering a defense case.[78]

In addition to pre-trial access to the evidence itself, and information about its assumptions and processes, other rights that are key to a meaningful ability to contest the results of digital and machine evidence include the ability to consult experts where necessary. David Sklansky has argued that a right to such expert testimony, and not merely in-court cross-examination, should be deemed a central part of the Sixth Amendment right of confrontation.[79]

The importance of a right of contestation in the algorithmic design process might be less obvious. But in a changing world in which machine evidence is not easily scrutinized at the trial itself, the adversarialism upon which common law systems are built might need to partially shift from the trial stage to the design and development stage. Carl DiSalvo has coined the term "adversarial design"[80] to refer to design processes that incorporate political contestation among different stakeholders. While adversarial design would not be a case-specific process, it could still involve representatives from the defense community. Others have suggested appointing a "defender general" in each jurisdiction[81] who could inject adversarial scrutiny into various recurring criminal justice issues at the front end. Perhaps such a representative could oversee defense involvement in the design, testing, and validation of algorithms. This process would supplement, not supplant, case-specific machine access and discovery.

OECD, Council on Artificial Intelligence, *Recommendation of the Council on Artificial Intelligence*, 2020, OECD/LEGAL/0449, at s. 1.3.iv, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[78] See *Holmes* v. *South Carolina*, 547 U.S. 319 (2006).

[79] See David A. Sklansky, "Hearsay's Last Hurrah" (2009) 2009:1 *Supreme Court Review* 1.

[80] Carl DiSalvo, *Adversarial Design* (Cambridge, MA: The MIT Press, 2012).

[81] See Daniel Epps & William Ortman, "The Defender General" (2020) 168:6 *University of Pennsylvania Law Review* 1469.

The right of contestation with regard to sophisticated AI systems, the methods of which may well never be meaningfully understood by the parties, might also need to incorporate a right to delegated contestation, in the form of the right to another machine's scrutiny of the results. Other scholars have noted the possibility of "reversible" algorithms that would audit themselves or each other,[82] or have suggested that one machine opinion alone should be deemed legally insufficient for a conviction, in the absence of corroboration from a second expert system.[83]

At the trial itself, the right of contestation should first include the right to argue for exclusion of the evidence on reliability (*Frye*/*Daubert*) and/or authenticity grounds. In the US federal system, proponents of digital and machine evidence must present sufficient evidence to persuade the factfinder that the evidence is what the proponent says it is, e.g., that an email is from a particular sender.[84] In China, courts have used blockchain technology to facilitate authentication of electronically stored information.[85] Jurisdictions' authenticity method might reasonably change as the ability for malfeasors to falsify evidence changes in the future. Likewise, litigants should have the right to insist on exclusion of machine evidence if inputs are not proven accurate. For example, in the United Kingdom, a "representation" that is made "other than by a person" but that "depends for its accuracy on information supplied (directly or indirectly) by a person" is not admissible in criminal cases without proof that the "information was accurate."[86] In some cases, this showing will require testimony from the inputter.[87]

[82] See Matthias Möller & Cornelis Vuik, "On the Impact of Quantum Computing Technology on Future Developments in High-Performance Scientific Computing" (2017) 19:4 *Ethics and Information Technology* 253.

[83] See "Machine Testimony", note 26 above, at 2038.

[84] See Federal Rules of Evidence, note 11 above, Rule 901(9) (allowing admission of a live witness to prove that a "process or system" produces an accurate result), and Rule 902(13), (14) (allowing admission of electronically stored and generated information upon presentation of a certification from a qualified witness who can attest to how the process works).

[85] See e.g. Zhuhao Wang, "China's E-Justice Revolution" (2021) 105:1 *Judicature* 37 (noting how blockchain is used for authentication of electronic evidence); Ran Wang, "Legal Technology in Contemporary USA & China" (2020) 39:10549 *Computer Law & Security Review* 1 at 4.

[86] Criminal Justice Act 2003, United Kingdom, c. 44, s. 129(1). If the inputter's "purpose" is "to cause … a machine to operate on the basis that the matter is as stated," it is treated as hearsay (see s. 115(3)), requiring the live testimony of the inputter (see s. 114(1)). The provision "does not affect the operation of the presumption that a mechanical device has been properly set or calibrated" (see s. 129(2)).

[87] See e.g. ibid. (requiring inputter testimony); Gert Petrus van Tonder, "The Admissibility and Evidential Weight of Electronic Evidence in South African Legal Proceedings:

Principle IV: Criminal defendants should have a right to a factfinding process that is epistemically competent but that retains a human in the loop, so that significant decisions affecting their liberty are not entirely automated.

Principle IV(a): While parts of the criminal process can be automated, human safety valves must be incorporated into the process to ensure a role for equity, mercy, and human moral judgment.

Both substantive criminal law and criminal procedure in the United States have become more mechanical over the past few decades, from mandatory arrest laws, to sentencing guidelines, to laws criminalizing certain quantities of alcohol in drivers' blood.[88] The more mechanical that the system becomes on the front end via, e.g., mandatory arrest, prosecution, liability rules, and sentencing, the more that safety valves such as prosecutorial, fact-finder, and sentencing discretion become critical to avoid inequities, i.e., results that are legal but unjust.[89] Moreover, mechanical regimes reduce the possibility of mercy, understood to mean leniency or grace, beyond what a defendant justly deserves. While mercy may be irrational, it is a pedigreed and "important moral virtue" that shows compassion and a shared humanity.[90]

As digital and machine evidence accelerate the mechanization of justice, jurisdictions should ensure that human actors are still able to exercise equity and mercy at the charging, guilt, and/or punishment stages of criminal justice. Not only are humans needed to ensure that laws are not applied mechanically. They are needed because they are literally human – they bring a human component to moral judgment that is necessary, if not for dignity, then at least for public legitimacy[91] and, in turn, for

A Comparative Perspective" (LLM thesis, University of Western Cape, May 2013), etd.uwc.ac.za/xmlui/bitstream/handle/11394/4833/VanTonder_gp_llm_law_2013.pdf (requiring live testimony of signer of documents).

[88] See generally Andrea Roth, "Trial by Machine" (2016) 104:5 *Georgetown Law Journal* 1245 ["Trial by Machine"] (noting how various aspects of American criminal justice have become more mechanical).

[89] See e.g. Martha C. Nussbaum, "Equity and Mercy" (1993) 22:2 *Philosophy & Public Affairs* 83 at 93 and n. 19 (explaining that equity "may be regarded as a 'correcting' and 'completing' of legal justice").

[90] Jeffrie G. Murphy, "Mercy and Legal Justice" in Jeffrie G. Murphy & Jean Hampton, *Forgiveness and Mercy* (Cambridge, UK: Cambridge University Press, 1998) 162 at 176.

[91] Meg Leta Jones, "Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood from Data Banks to Algorithms" (2017) 47:2 *Social Studies of Science* 216 at 231.

enforcement of criminal law.[92] In the United States, scholars have written since the 1970s of the illegitimacy of verdicts based solely on "naked statistical evidence," based on personhood concerns.[93] Moreover, humans add to the fact-finding process as well, rendering AI systems fairer without having to make such systems less accurate through simplification.[94] Corroborating these observations, recent AI guidelines and data privacy laws reflect the public's desire to keep humans in the loop with regard to automated decision-making, from the Council of Europe's call to "ensure that the introduction, operation and use of AI applications can be subject to effective judicial review,"[95] to the EU Directive prohibiting processes that produce an "adverse legal effect" on a subject "based solely on automated processing," without appropriate "safeguards for the rights and freedoms of the data subject, at least the right to obtain human intervention on the part of the controller."[96]

More concretely, criminal liability should not be based solely on an automated decision. Red light cameras are the closest the United States has come to fully automated liability, but thus far, such violations end only in a mailed traffic ticket rather than a criminal record. Moreover, in jurisdictions with juries, the power of jury nullification should continue undisturbed. It may well be that jurors' ability to decide historical fact, e.g., "was the light red?", could be curtailed, so long as their ability to decide evaluative data, e.g., "did the defendant drive 'recklessly'?", is preserved.[97] Indeed, some historical fact-finding might be removed from lay jurors, if they lack the "epistemic competence" to assess the evidence's probative value.[98]

[92] See generally Tom Tyler, "Procedural Justice, Legitimacy, and the Effective Rule of Law" (2003) 30:1 *Crime & Justice* 283 (explaining the role of procedural justice in inspiring compliance with law).

[93] See e.g. Laurence Tribe, "Trial by Mathematics: Precision and Ritual in the Legal Process" (1971) 84:6 *Harvard Law Review* 1329.

[94] See e.g. Katharine Miller, "When Algorithmic Fairness Fixes Fail: The Case for Keeping Humans in the Loop," *Stanford University: Institute for Human-Centered AI* (November 2, 2020), https://hai.stanford.edu/blog/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop.

[95] "Justice by Algorithm", note 47 above, at 9.13.

[96] See European Commission, Directive (EU) 2016/680 of April 27, 2016 (OJ 4.5.2018, L 119, 89), Art. 11.

[97] Others have called for this; see e.g. Josh Bowers, "Legal Guilt, Normative Innocence, and the Equitable Decision Not to Prosecute" (2010) 110:7 *Columbia Law Review* 1655 at 1723; Anna Roberts, "Dismissals as Justice" (2017) 69:2 *Alabama Law Review* 327 (discussing Model Penal Code §2.12).

[98] See e.g. Scott Brewer, "Scientific Expert Testimony and Intellectual Due Process" (1998) 107:6 *Yale Law Journal* 1535 at 1551 (arguing for a due process right to an "epistemically competent" fact-finder).

Jurisdictions could still ensure that humans remain in the loop by disallowing machine experts from giving dispositive testimony on ultimate questions of fact,[99] prohibiting detention decisions based solely on a risk assessment tool's score, and requiring a human expert potentially liable for injustices caused by inaccuracies to vouch for the results of any machine expert, before introducing results in a criminal proceeding.

Principle IV(b): Jurisdictions should ensure against automation complacency by developing effective human–machine interaction tools.

Keeping a human in the loop would be useless if that human deferred blindly to a machine. For example, if sentencing judges merely rubber-stamped scores of risk assessment tools, there would be little reason to ensure that judges remain in the loop.[100] Likewise, if left to their own devices, juries might irrationally defer to the apparent objectivity of machines.[101] A human in the loop requirement should entail the development of tools to guard against automation complacency. One underused tool in this regard is jury instructions. For example, where photographs are admitted as silent witnesses, the jury hears little about lens, angle, speed, placement, camera-person bias, or other variables that might lead it to draw a false inference from the evidence. The jury should be educated about the effect of these variables on the image they are assessing.[102] Ultimately, jurisdictions should draw from the fields of human factors engineering, and human–computer interaction and collaboration, in designing ways to ensure a systems approach that keeps humans in the loop while leveraging the advantages of AI.

Principle IV(c): Jurisdictions should establish a formal means for stakeholders to challenge uses of digital and machine evidence that are fundamentally inconsistent with principles of human-delivered justice.

[99] Cf. Federal Rules of Evidence, note 11 above, Rule 704 (prohibiting expert witnesses from giving opinions as to whether criminal defendants have the mental state required).

[100] See Sonja B. Starr, "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination" (2014) 66:4 *Stanford Law Review* 803 at 866–868 (suggesting that actuarial instruments drive judicial sentencing decisions).

[101] R. A. Bain, "Comment, Guidelines for the Admissibility of Evidence Generated by Computer for Purposes of Litigation" (1982) 15:4 *UC Davis Law Review* 951 at 961 (noting that fact-finders might be unduly "awed by computer technology").

[102] See Benjamin V. Madison III, "Seeing Can Be Deceiving: Photographic Evidence in a Visual Age – How Much Weight Does It Deserve?" (1984) 25:4 *William & Mary Law Review* 705 at 740 (arguing for jury instructions along these lines for photographs); see generally Jessica M. Silbey, "Judges as Film Critics: New Approaches to Filmic Evidence" (2004) 37:2 *University of Michigan Journal of Law Reform* 493 (suggesting trial safeguards for explaining testimonial infirmities of images to fact-finders).

Keeping a human in the loop also necessarily means taking steps to ensure against inappropriate uses of AI that threaten softer systemic values like dignity. For example, certain machines might be condemned as inherently dehumanizing, such as the penile plethysmograph[103] or deception detection.[104] Just as some modes of obtaining evidence are rejected as violating substantive due process, such as forcibly pumping a suspect's stomach to find evidence of drug use,[105] modes of determining guilt should be rejected if the public views them as inhumane. Other jurisdictions might decide that the "right to explanation" is so critical to public legitimacy that overly complex AI systems must be abandoned in criminal adjudication, even if such systems promise more accuracy.[106] Whatever approach jurisdictions adopt regarding these issues, they should resolve such issues first, and only then look for available technological enhancements of proof, rather than vice versa. Numerous scholars have written about the seduction of quantification and measurement,[107] and the Council of Europe expressly included in its guidelines for the use of AI in criminal justice that Member States should "ensure that AI serves overall policy goals, and that policy goals are not limited to areas where AI can be applied."[108]

III Conclusion

The principles for governing digital and machine evidence articulated in this chapter attempt to move beyond the adversarial/inquisitorial divide, and incorporate the thoughtful recent work of so many scholars, policymakers, and stakeholders worldwide in promulgating guidelines for the ethical and benevolent use of AI in decision-making affecting peoples'

[103] "Trial by Machine", note 88 above (describing the penile plethysmograph and arguing that its use violates dignitary interests of subjects).

[104] See ibid. (discussing personhood objections to various forms of lie detection evidence).

[105] See *Rochin* v. *California*, 342 U.S. 165 (1952).

[106] See e.g. "Justice by Algorithm", note 47 above, at 9.9 (Member States should "ensure that the essential decision-making processes of AI applications are explicable to their users and those affected by their operation").

[107] See e.g. Andrea Saltelli, "Ethics of Quantification or Quantification of Ethics?" (2020) 116:102509 *Futures* 1 (discussing "metric fixation"); "Trial by Machine", note 88 above, at 1281 (quoting Sally Engle Merry, *The Seductions of Quantification: Measuring Human Rights, Gender Violence and Sex Trafficking* (Chicago, IL: University of Chicago Press, 2016) (exploring the distorting effects of the quest for measurable indicators in the context of human rights)).

[108] "Justice by Algorithm", note 47 above, at 9.3.

lives. Applied to a common law adversarial criminal system such as that in the United States, these principles may manifest in existing statutory and constitutional rights, albeit in new ways. Applied to other nations' systems, these principles will manifest differently, perhaps because such systems already recognize the need for "out of court evidence gathering"[109] to ensure meaningful evaluation of complex evidence. On the other hand, as Sabine Gless has suggested, continental systems might find that party-driven examinations have an underappreciated role to play in ensuring reliability of machine evidence.[110]

As AI becomes more sophisticated, one key goal for all justice systems will be to ensure that AI is not merely given an objective to accomplish, such as "determine whether this witness is lying" or "determine if this person contributed to this DNA mixture," but is programmed to continually look to humans to express and update their preferences. If the former occurs, AI will preserve itself at all costs, and may engage in behavior antithetical to human values, to get there.[111] Only if machines are taught to continually seek feedback can AI remain benevolent. We cannot simply program machines to achieve the goals of criminal justice – public safety, social cohesion, equity, the punishment of the morally deserving, and the vindication of victims. We will have to ensure that humans have the last word on what justice means and how to achieve it.

[109] "AI in the Courtroom", note 5 above, at 251.
[110] "AI in the Courtroom", note 5 above, at 249.
[111] See generally Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York, NY: Penguin Books, 2019).