Sofia Bonicalzi

A matter of justice. The opacity of algorithmic decision-making and the trade-off between uniformity and discretion in legal applications of artificial intelligence

1. *Introduction*

In the last few years, decisions about matters of distributive (concerning resource allocation) and retributive (concerning the punishment of lawbreakers) justice have been more and more outsourced to automated systems (A.I.), and unprecedented ethical challenges have progressively emerged. In the realm of retributive justice, the usage of A.I., usually limited to the pre-trial and post-trial phase, ranges from individuating criminals to providing companionship for inmates and allocating cases to specific judges. In the field of distributive justice, A.I. is involved, for instance, in decisions about social housing, access to health care, or career promotions (Jorgensen 2022; Rai 2020; Završnik 2020).

As compared to human adjudicators, A.I. presents, or may present in the future, concrete advantages in terms of efficiency (e.g., time and cost reduction) and uniformity of performance. However, its contribution to legal decision-making must be carefully assessed given its potential ethical drawbacks and impact on basic human rights, such as the right to be tried before an independent tribunal and to access a human rights-based criminal justice system, the presumption of innocence, the respect of privacy, or the right to equal access to public goods and services[1]. This is particularly

[1] For instance, in the U.S. legal system, the requirement that a state governs impartially, and grants people equal protection is imposed by the Equal Protection clause of the Fourteenth

DOI: 10.4454/teoria.v42i2.161

so since current A.I. systems are progressively moving from auxiliary tools to primary decision-makers, able to impact more directly on people's life by taking decisions and making recommendations and predictions[2].

This paper aims to discuss a specific challenge – the difficult trade-off between uniformity and discretion in judicial applications of A.I. – against the backdrop of current debates in philosophy, cognitive science, and artificial intelligence. This is the gist of it: the usage of A.I. has been notoriously criticized with reference to the so-called *black box problem*, which arises in virtue of the lack of transparency and interpretability characterizing algorithmic decision-making, especially when developed through unconstrained or unsupervised deep learning. Emphasis has thus been placed on how to make the procedures more transparent through forms of explainable A.I. (§ 2). That said, it would be myopic to assume that humans alone are *de facto* best reasoners and transparent deliberators. Cognitive sciences have indeed widely shown that human reasoning is affected by multiple cognitive limitations and biases that might be likewise detrimental to the equity and fairness of the judicial processes. It is not by chance that A.I.-based products are marketed not just as up to mimicking human intelligence but also as in principle able to do better than humans by overcoming common cognitive fragilities. The implementation of A.I. technologies would thus promote a general increase in the uniformity of both procedures and outcomes and cut down pernicious excesses of discretion. A standard reply of those who warn against the potentially despicable effect of A.I. on judicial standards is that this goal is currently out of reach. Indeed, A.I.-based systems are not immune from biases analogous to those that they promise to tame. These anomalies are usually tied to the inclusion of such biases in the set of data through which the algorithm has been trained, to the *under-representativity* of the data, or to wrong assumptions involuntarily made by the programmers (§ 3).

Amendment to the Constitution (Biddle 2020). With expressions like "legal decision-making" and "judicial ruling", here I will refer in general to the activities of the various agencies dealing with distributive and retributive justice broadly conceived. Further work would be needed to discuss in detail the applications of A.I.-based systems to these different domains.

[2] As compared to previously existing supporting tools, «an AI system is a machine-based system that makes recommendations, predictions or decisions for a given set of objectives. It does so by: (i) utilising machine and/or human-based inputs to perceive real and/or virtual environments; (ii) abstracting such perceptions into models manually or automatically; and (iii) deriving outcomes from these models, whether by human or automated means, in the form of recommendations, predictions or decisions» (*Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights*, 2019).

In bracketing the issue of *algorithmic bias*, here I focus on a different argumentative line that stresses the positive value of flexibility and discretion, emphasizing that sidestepping the peculiarities of human reasoning might even have some detrimental effects on the fairness of justice administration. This is particularly the case when this process is conducive to the elimination of reasonable standards of flexibility and discretion, including the ability to bend the rules when circumstances so require. Consequently, the contribution of both human and automated decision-making to social justice matters must be carefully balanced if fair results are to be obtained (§ 4)[3].

2. *The opacity of algorithmic decision-making*

A notorious problem concerning the workings of A.I. is known as the *black box problem*. This refers to the incomprehensibility or lack of transparency regarding how the system moves from the provided inputs to the produced outputs. In some situations, this opacity might be due to contingent issues, such as when the process is protected by trade secret law so that the defendants are not granted a meaningful explanation of the output. A widely debated, and infamous, case of this kind is *Loomis v. Wisconsin.* The defendant (Loomis) – categorized at high risk for recidivism and sentenced to six years in prison and five years of supervision – was denied access to the procedures and methodologies through which the predictive algorithm COMPAS (*Correctional offender management profiling for alternative sanctions*) issued the relevant risk assessment report[4]. This is especially problematic in non-autocratic systems valuing the distribution of decision-making power. Within democratic systems, the defendants, who already find themselves in a vulnerable position, tend to be

[3] Whereas the examples I will refer to belong to the common law tradition and literature, judicial discretion is central even in the systems adopting civil law, whenever general principles must be adapted to specific circumstances. Comparing the role of judges in common law and civil law systems, Yu (1999) suggests that, despite profound differences remain, there has been a progressive convergence between the two so that even «the courts in the civil law countries perform a law-making function through an extension by a flexible process of interpretation and by express legislative instruction. Judicial adaptation to changing circumstances is facilitated by the so-called "general clauses" which leave lots of discretion to the judge»

[4] Loomis appealed the sentencing for violating his *due process rights* (protected, in the U.S. systems, by the Fifth and the Fourteenth Amendment) for various reasons, including that the opacity of the algorithms prevented him from assessing its accuracy and that his right to an individualized sentence was violated (Biddle 2020).

seen as entitled to a comprehensible explanation for them to actively participate in the process and eventually question its output (Završnik 2020).

In other cases, this opacity is more radical to the extent that even programmers themselves may find the process difficult or impossible to interpret or verify, independently of whether the procedures are protected or publicly accessible. On the technical side, various causes, in isolation or combination, may explain this lack of transparency. For instance, the opacity might be related to the system's combining multiple variables that exceed standard cognitive abilities. Or it might depend on the system's relying on complex correlations, statistical modeling, or deep learning techniques that are irreducible to logical reasoning and argumentation (Re and Solow-Niederman 2019). Moreover, technical inexplicability in a descriptive sense may or may not underlie the normative indefensibility of a decision, thus creating a further level of opacity: when descriptive or technical explanations of how a decision has been made are inaccessible, it might become difficult to assess whether the ensuing normative evaluation is fair at all (Selbst and Barocas 2018).

A moment of caution is warranted to the extent that opacity is not limited to A.I. adjudicators. Traditional judicial procedures may also involve some hidden variables, such as when adjudicators rely on anonymous witnesses and undisclosed documentary evidence to make their decisions (see Binger v. King Pest Control). However, the use of undisclosed evidence or witness testimony is usually limited as much as possible and must comply with strict regulatory norms rather than being considered an almost unavoidable part of the adjudication process (Završnik 2020).

Furthermore, the kind of opacity that is typical of automated systems, where procedures are systematically subtracted from public oversight, has been criticized for bringing about a higher risk of alienation of both laypeople and experts. Indeed, a common (although sometimes misleading) assumption is that traditional judicial decisions to some extent mirror human reasoning and its everyday logic. As such, they can be discussed or even challenged via standard argumentative strategies that are within the reach of both experts and novices. From a sociological angle, a radical lack of understanding, especially affecting those who do not have the appropriate technical background, implies vulnerability to the law procedures and generates power imbalances (Re and Solow-Niederman 2019).

From a psychological angle, this issue can be understood as part of the more general discussion about how human reasoning is distinctively affected by the interaction with A.I. and technologies in general. The reliance on A.I. and other forms of automated systems to carry out daily tasks, inside and outside the legal system, is itself associated with a host of automation-induced distortions in cognition and performance, including skill degradation, automation complacency, and automation bias. Skill degradation refers to the loss of skills that might occur due to the outsourcing of some activities to A.I.[5], which often goes together with phenomena such as automation complacency (the uncritical, passive, and potentially diffident reliance on technologies that are seemingly more competent than the user) and automation bias (the preference for automated solutions that are seen as more reliable than human-based solutions in cases of mismatches or antithetical information) (Parasuraman and Manzey 2010).

Therefore, while inscrutability per se may lead to a general rejection of A.I. adjudicators (Rai 2020), overtrust or loafing may represent the opposite end of the spectrum (Zerilli, Bhatt, and Weller 2022), and even promote a worrisome self-reinforcing circle: the surge in the implementation of A.I. systems may boost skepticism about the practices and competencies of traditional human adjudicators, which may, in turn, increase the social pressure to turn to A.I. solutions to societal problems.

In the last few years, there has been more and more emphasis on how to make the procedures accessible through forms of interpretable or explainable A.I. The goal is to develop methods and processes that can be better understood, and potentially controlled, by humans[6]. On the technical side, the demand for transparency is not easy to address, especially because the improvement in functionality is often afforded by a corresponding increase in complexity. On the theoretical side, this demand requires further articulation, for instance in terms of specifying the level of understanding that is required, the content that must be communicated to different stakeholders, or the amount of information that allows subjects to make more informed decisions without exposing them to an excessive burden (Biddle 2020). Moreover, it is crucial to notice that the effort to generate understandable elucidations may produce fake but intuitive explanations that superficially satisfy the human need for a narrative but are not representative of the

[5] Research has shown that the impact of A.I. on work-related skills may vary depending on the type of job, with a potential increase in skills in high-skill jobs and an opposite effect on low-skill ones (Holm and Lorenz 2022).

[6] For an overview of interpretable models see Hall and Gill 2019; Linardatos, Papastefanopoulos, and Kotsiantis 2020; Rai 2020.

underlying processes (Perez, 2018; Re and Solow-Niederman, 2019). As such, the problem of how to balance complexity and explainability remains an open task for future research.

3. *The fragilities of human reasoning in the legal setting and the algorithmic bias*

While lack of transparency remains an issue in human-computer interaction, cognitive sciences have widely shown that human reasoning can be opaque as well, potentially affecting the equity and fairness of the decision-making process. Decades of literature in cognitive and social psychology have consistently suggested that apparently rational decisions can be surreptitiously determined by automatic or unconscious processes shirking metacognitive reflection (Bargh and Chartland 1999). Cognitive biases of various sorts have been shown to affect professional adjudicators as well, despite their tendency to consider themselves extraordinarily resistant to them in virtue of their training and expertise – and despite their ability to appear unbiased and impartial (Edmond and Martire 2019).

A most emblematic example is provided by discussions about the spread of implicit biases among legal practitioners (Rachlinski and Johnson 2009), stimulating reflections about the role that social cognition research should play in the law (Borgida and Girvan 2015) and the need to take affirmative action to counter discrimination (Kang and Banaji 2006). Implicit biases in particular are tied to forms of unintentional discrimination reflecting problematic social stereotypes that have the potential to distort the ensuing judgment (Holroyd 2015). While decision-makers themselves may fall prey to such biases, the problem is even self-reinforcing in judicial cases that deal directly with discriminatory practices: despite ubiquitous information about the existence of implicit biases[7], the likelihood that adjudicators, even informed ones, will take the offenders' implicit biases seriously when judging their conduct remains low (Girvan, 2015).

In the legal context, cognitive biases of various sorts are particularly worrisome to the extent that the integrity of the legal processes depends on adjudicators' being independent and impartial (see Ebner v Official Trustee) and on their making decisions based exclusively on admissible evidence. As discussed by Edmond and Martire 2019, evidence shows that

[7] But see Macherie 2022 for an extended critique of the science of implicit attitudes.

adjudicators are affected by common cognitive biases, including anchoring effects whereby prior exposure to arbitrary numeric information affects subsequent high or low sentencing decisions (Englich, Mussweiler, and Strack 2006). Furthermore, they are demonstrably sensitive to expectancy effects, e.g., judges' beliefs about the defendants' culpability can be passed to the jurors via mechanisms of nonverbal communication, thus affecting the final verdict (Rosenthal 2003). The impact of contingent, supposedly irrelevant, factors in the judicial ruling has been highlighted by a debated study by Danziger and colleagues (2011), testing the popular saying according to which justice is "what the judge ate for breakfast". The study shows that the percentage of favorable rulings on a judge's typical working day gradually drops during a section of sequential decisions and is restored after the judge takes a break. The suggested explanation is that the ruling effort progressively depletes the judge's executive functions and mental resources so that she is more and more inclined to avoid intervening and to accept the *status quo*, in this case by rejecting the defendant's request.

Given such difficulties and lack of impartiality, it is not surprising that A.I. systems are presented as in principle able to provide valuable support to traditional judicial processes, by overcoming the disturbing variability that is typical of human decision-making. As compared to human decision-making, A.I. is more stable and efficient: it does not decline over time, it is not subjected to contingent environmental influences, and it is never tired. These features secure advantages in terms of reduction of time and costs as well as in terms of uniformity of performance.

A major issue, however, is that it is far from being clear to which extent the A.I. systems currently available in the legal industry can make impartial and unbiased decisions (Dressel and Farid 2018). The problem of algorithmic bias, now officially tackled by some discussed legislative initiatives[8], refers to systematic errors in computer-based systems producing unfair outcomes that benefit given social groups and thus reinforce existing stereotypes (Belenguer 2022). In this respect, a wide-ranging debate was sparked off by the empirical audit run in 2016 by ProPublica (a nonprofit New York-based organization dedicated to investigative journalism), reporting that the recidivism algorithm COMPAS, implemented by Equivant

[8] See, for instance, the report issued by Human Rights Watch (*How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers*) on the European proposed legislation on A.I.

and widely used in the U.S., was biased against black (Angwin, Larson, and Kirchner 2016. See also Dieterich, Mendoza, and Brennan 2016).

While the problem of algorithmic bias remains one of the main challenges for the future of human-A.I. interaction, here I will bracket this issue and focus on a different type of challenge. Indeed, even assuming that the problem of algorithmic bias can be solved, it is unclear that the neutrality and uniformity of results that A.I. can achieve are always positive. In § 4, I will thus discuss this challenge in terms of the trade-off between uniformity and discretion in judicial applications of A.I.

4. *The trade-off between uniformity and discretion in judicial applications of A.I.*

Stressing the alleged neutrality of A.I. adjudicators and the uniformity of their performance implicitly suggests that flexibility and discretion, which are more typical of human intelligence, must be curbed as much as possible. Excess of the bounds of discretion must indeed be condemned, especially so when they cross over into arbitrariness or discrimination. However, flexibility and discretion often play a positive societal role that should not be neglected, particularly when they are not detrimental to social groups that are socially dispossessed – avoiding disadvantaging those who are already disadvantaged is indeed a basic ethical principle that is accepted in most theories of justice (Biddle 2020).

If we look at the common law tradition, a historical distinction is that between *legal formalism* and *legal realism*. Legal formalism maintains that legal decision-making consists in the mechanical and logical application of legal rules and reasons. Conversely, legal realists hold that a host of psychological, contextual, and cultural factors shape the adjudicators' decisions so that they ultimately deliberate based on what seems fair in a given situation (Leiter 2005; Posner 1986). The spirit of the legal realist tradition – caricatured in the study by Danziger, Levav, and Avnaim-Pesso 2011 where the judges' ruling habits depend on when they have had breakfast – is expressed by Oliver Wendell Holmes' famous saying, according to which «the life of the law has not been logic; it has been experience» (Holmes 1881/1991). Whereas this is a descriptive claim about how judicial ruling unfolds in practice, it acquires a normative force to the extent that fairness in adjudication can be often achieved precisely through the exercise of reasonable discretion.

While A.I. systems may achieve better results in terms of uniformity of performance, a specific challenge is thus how to preserve the appropriate amount of discretion in judicial ruling. In this respect, legal scholars have observed that A.I. may affect not just the modalities of legal decision-making (e.g., in terms of time and cost reduction) but the very same values that inform and shape the existing legal culture: «by offering efficiency and at least an appearance of impartiality, AI adjudication will foster a turn toward 'codified justice', that is, a paradigm of adjudication that favors standardization above discretion» (Re and Solow-Niederman 2019, 246).

From a psychological point of view, cognitive sciences have shown that people's subjective sense of distributive and retributive justice cannot be reduced to the mechanical applications of pre-existing rules. More specifically, it cannot prescind from the flexible integration of multiple processes, including both rational and emotional factors. This is evident, for instance, in economic decision-making about equity in resource distribution. In playing the ultimatum game, people notoriously violate the standard norms of rationality when rejecting unfair offers or acting spitefully (Pillutla and Murnighan 1996). Unfair offers are known to generate a conflict between competing tendencies: the emotional one, mediated by the bilateral anterior insula, consists in resisting the offer, while the rational one, mediated by the dorsolateral prefrontal cortex, consists in accepting the offer (Sanfey et al. 2003).

Furthermore, studies about evaluations of procedural and distributive unfairness in resource allocation show a marked dissociation of activation between the two types of judgments, with unfair procedures eliciting greater activation in brain areas concerned with social cognition (ventrolateral prefrontal cortex, superior temporal sulcus) and unfair outcomes eliciting greater activation in areas involved in emotional processing (anterior cingulate cortex, anterior insula, dorsolateral prefrontal cortex) (Dulebohn at al. 2009). Analogously, the subjective sense of retributive justice in third-party scenarios results from the combination of the affective evaluation (in the amygdala, medial prefrontal, and posterior cingulate cortex) of crime severity (how much should the offender be punished?) and the more rational and categorical evaluation (in the dorsolateral prefrontal cortex) of individual responsibility (is the offender responsible or not?) (Buckholtz et al. 2008).

The involvement of affective processes in the evaluation of various forms of distributive and retributive justice should not be viewed simplistically as the result of cognitive biases and distortions affecting human rationality. Conversely, these psychological mechanisms are fundamental to the formation of our subjective sense of justice and, at least according to the realist tradition, also to how justice works (and perhaps should work) in practice, i.e., in ways that are *responsive* to the agenthood of the person who is being judged rather than being based merely on abstract inferences about one's social identity (Dworkin 1977; Jorgensen 2022).

Therefore, we are now able to better understand what difficult trade-off is at stake: on the one hand, the possibility to get rid of the *bad* variability that is typical of human cognition is extremely valuable – providing that A.I. can overcome the vexed issues of opacity and algorithmic bias. On the other hand, one should not throw the baby out with the bathwater, i.e., one should avoid giving up the good flexibility and discretion that, in humans, depend both on the joint work of different cognitive processes (emotional and rational) and on the flexible adaptation to cultural changes and specific circumstances.

In discussing this challenge, Re and Solow-Niederman 2019 note that the dichotomy between flexible humans and fixed A.I. systems should not be exaggerated[9]: it is contingent and not unavoidable, and, in any case, possible solutions are not free from specific difficulties[10]. For instance, one fascinating option could be that of coding the ability for discretion directly into the A.I. system. This kind of programmable flexibility should reflect the social, moral, and legal consensus on a given topic and evolve in relation to societal changes or unanticipated tasks. This solution, however, poses problems both at a technical and normative level insofar as it is unclear how and when this flexibility can (or should) be implemented, and to what extent forms of good and bad discretion can be so neatly disentangled – this even if one brackets reasonable concerns about the limits of an opaque, black-box based, exercise of discretion.

Alternatively, one may hypothesize a sort of division of labor between human adjudicators and A.I., to be implemented via collaborations in different phases of the judicial process relative to the same cases or by restricting the usage of A.I. to selected cases. Working in tandem with A.I., human adjudicators will play a major role whenever the ability to exert dis-

[9] For instance, in virtue of its ability to consider a higher number of variables, the A.I. could be even more sensitive than humans to the nuances of the situation.

[10] Furthermore, it might be the case that personalization leads to «being treated *worse* than otherwise and is in some tension with other weighty principles of justices, such as the generality and equal application of law, and the fair social distribution of various burdens» (Jorgensen 2022).

cretion looks particularly valuable. However, finding ways to determine the appropriate equilibrium between human and A.I. adjudicators is not simple: the situations in which one wants to avoid bad (i.e., biased) discretion are often the same in which positive (i.e., attuned with the specific situation) discretion is to be praised.

On the practical side, the mechanical and standardized application of existing rules may cause troubles particularly affecting the socially dispossessed, and further exacerbate the very same forms of discrimination it is supposed to fight. Some examples of this, in the domain of both retributive and distributive justice, can be found in Virginia Eubanks' book *Automating Inequality* (2018). Concerning retributive justice, Eubanks quoted a 2000 report of the *Leadership Conference on Civil and Human Rights* taking stock of several *mandatory minimum sentencing laws* enacted in the U.S. in the previous decades and limiting the adjudicators' discretion. Based on the acquired evidence in terms of racial disparity in the outcomes of the criminal justice system, the report states explicitly that «minorities fare much worse under mandatory sentencing laws and guidelines than they did under a system favoring judicial discretion. By depriving judges of the ultimate authority to impose just sentences, mandatory sentencing laws and guidelines put sentencing on auto-pilot». This is paradoxical to the extent that one of the main justifications for automatizing justice is usually to reduce imbalances in the treatment of different social groups.

Regarding distributive justice, Eubanks discusses the impact of algorithms used to allocate social housing based on risk profiles. An emblematic case study is offered by the social housing program *Home for Good*, implemented in 2013 to fight homelessness in the run-down area of Skid Row (Los Angeles). Having the potential to simplify pre-existing processes and potentially limit the impact of the providers' implicit bias, the system was implemented through an assessment tool that collected information and ranked the homeless based on their vulnerability score. Eubanks recognizes that the program successfully managed to help several people with a history of unstable housing. However, some of the interviewed people reported that they were automatically denied help and that the system acted as a black box since no explanation was usually provided about their prioritization score. Moreover, some major issues persist in the way the algorithms were used to track and monitor the poor: people whose behavior and lifestyle were classified (based on the Vulnerability Index Tool VI-SP-DAT) as particularly risky or even illegal scored higher on the priority list while being at the same time subjected to higher scrutiny and potentially face jail time. As such, the program is not just a tool to match the homeless to the housing resources, but a surveillance system aiming to control and criminalize the socially dispossessed while lacking the individualized attention and the ability to bend the rules that were typical of older forms of surveillance and assistance.

To conclude, when judging the functioning of A.I. systems, one should be careful not to confuse different forms of *impartiality*: one thing is to say that algorithms are (potentially) less biased (and thus more impartial) than human adjudicators towards specific social groups. A quite different thing is to wrongly assume that A.I. based decisions can then be automatically impartial also with respect to given ethical, social, and political models and values. Conversely, algorithms remain value-laden, although in ways that, once again, might remain opaque to individual citizens: its functioning reflects a specific trade-off between different interests and values, and ultimately between different societal models and conceptions of fairness, e.g., about how social cooperation must be organized. These differences are embedded, for instance, in the epistemic risks and failures the system is set to tolerate when making decisions about resource allocation or appropriateness of a certain treatment, in the judgments about what problems must be prioritized and require intervention, or in how single factors will weigh in on judicial decisions (e.g., the extent to which a person's socioeconomic background must be considered by risk assessment tools) (Biddle 2020). Therefore, as much as with human adjudicators and laws, relying on A.I. systems requires addressing ethical, and not just technical, difficulties about the societal models that are to be implemented.

5. *Conclusion*

In this paper, I have briefly discussed some ethical challenges linked with the implementation of A.I. systems in judicial decision-making. A.I. seemingly guarantees advantages in terms of uniformity and efficiency of performance, potentially overcoming the typical biases and variability of human adjudicators. However, even assuming that the problem of algorithmic bias can be somehow addressed in the future (and that reasonable choices about the overarching values embedded in A.I. decisional procedures are made), the progressive automatization of the justice system may bring along the neutralization of the good forms of flexibility and discretion that are more typical of human intelligence. This flexibility and discretion have to do with the agent's *right to be treated as an individual*, which goes beyond the right to be judged by an unbiased adjudicator and rather concerns the adjudicator's «duty to be responsive to the individual's responsible agency» or to respect the separateness of persons (Jorgensen 2022). Especially if one is already in a vulnerable position, being treated as an individual includes the right to be part of the decision-making process, which might be further eroded by the opacity of the algorithmic procedures. As such, finding the appropriate balance between uniformity and discretion appears to be one of the major challenges in judicial applications of human-A.I. interaction.

Acknowledgments

I thank the members of the audience of the workshop "Nuove sfide nei processi di decisione. Bioetica, Neuroetica, Etica dell'Intelligenza Artificiale" at the University of Pisa for their insightful comments on an early version of this paper. I am also grateful to an anonymous reviewer for his/her suggestions.

Funding

S.B. has benefitted from the PRIN 20175YZ855.

References

Angwin, J., Larson, J., Kirchner, L. *Machine Bias*. *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bargh, J.A., and Chartland, T.L. "The Unbearable Automaticity of Being." *American Psychologist* 54 (1999): 462-479.

Belenguer, L. "AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry." *AI Ethics* (2022): 1-17.

Biddle, J. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." *Canadian Journal of Philosophy*, (2020): 1-21.

Binger v. King Pest Control, 401 So. 2d 1310 (Fla. 1981).

Borgida, E., and Girvan, E.J. "Social Cognition in Law." In *APA Handbook of Personality and Social Psychology*, *vol. 1*, *Attitudes and Social Cognition*, edited by M. Mikulincer, P.R. Shaver, E. Borgida, and J.A. Bargh, 753-774. *American Psychological Association*, 2015.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., and Marois, R. "The Neural Correlates of Third-Party Punishment." *Neuron* 60, no. 5 (2008): 930-940.

Danziger, S., Levav, J., and Avnaim-Pesso, L. "Extraneous Factors in Judicial Decisions." *PNAS* 108, no. 17 (2011), 6889-6892.

Dieterich, W., Mendoza, C., and Brennan, T. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Performance of the COMPAS Risk Scales in Broward County*, Northpointe Inc. Research Department. July 8, 2016. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Dressel, J., and Farid, H. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Sci Adv.*, 4, no. 1 (2018): eaao5580.

Dulebohn, J.H., Conlon, D.E., Sarinopoulos, I., Davison, R.B., and McNamara, G. "The Biological Bases of Unfairness: Neuroimaging Evidence for the Distinctiveness of Procedural and Distributive Justice". *Organizational Behavior and Human Decision Processes* 110, no. 2 (2009): 140-151.

Dworkin, R. *Taking Rights Seriously*. London: Duckworth, 1977.

Ebner v Official Trustee in Bankruptcy, 205 CLR 337; 2000 HCA 63.

Edmond, G., and Martire, K.A. "Just Cognition: Scientific Research on Bias and Some Implications for Legal Procedure and Decision-Making." *The Modern Law Review* 82, no. 4 (2019): 633-664.

Englich, B., Mussweiler, T., and Strack, F. "Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making." *Personality and Social Psychology Bulletin* 32, no. 2 (2006): 188-200

Eubanks, V. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York, NY: St Martins Pr., 2018.

Girvan, J.E. "On Using the Psychological Science of Implicit Bias to Advance Anti-Discrimination Law." *George Mason University Civil Rights Law Journal* 26, no. 3 (2015).

Hall, P., and Gill, N. *An Introduction to Machine Learning Interpretability*. Second Edition. Sebastopol, CA: O'Reilly Media, Incorporated, 2019.

Holm, J.R., and Lorenz, E. "The Impact of Artificial Intelligence on Skills at Work in Denmark." *New Technology, Work and Employment* 37, no. 1 (2022): 79-101.

Holmes, O.W. *The Common Law*. Mineola, NY: Dover Publications, 1881/1991.

Holroyd, J. “Implicit Bias, Awareness and Imperfect Cognitions.” *Consciousness and Cognition* 33 (2015): 511-523.

“How the EU’s Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers,” Human Rights Watch, last modified November 10, 2021, https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net#_ftn64.

Jorgensen, R. “Algorithms and the Individual in Criminal Law.” *Canadian Journal of Philosophy* 52, no. 1 (2022): 61-77.

Kang, J., and Banaji, M.R. “Fair Measures: A Behavioral Realist Revision of ‘Affirmative Action.” *California Law Review* 94, no. 4 (2006): 1063-1118.

Leiter, B. “American Legal Realism.” In *The Blackwell Guide to Philosophy of Law and Legal Theory*, edited by W. Edmundson and M. Golding, 50-66. Oxford: Blackwell, 2005.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. “Explainable AI: A Review of Machine Learning Interpretability Methods.” *Entropy* 23, no. 1 (2020): 18.

Macherie, E. “Anomalies in Implicit Attitudes Research.” *Wires Cognitive Science* 13, no. 1 (2022): e1569.

Parasuraman, R., and Manzey, D.H. “Complacency and Bias in Human Use of Automation: An Attentional Integration.” *Hum Factors* 52, no. 3 (2010): 381-410.

Perez, C.E. *Deep Learning’s Uncertainty Principle*. April 6, 2018. https://medium.com/intuitionmachine/deep-learnings-uncertainty-principle-13f3ffdd15ce#:~:text=The%20uncertainty%20principle%20as%20applied,interpretable%20don%27t%20generalize%20well.

Pillutla, M.M., and Murnighan, J.K. “Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers.” *Organizational Behavior and Human Decision Processes* 68, no. 3 (1996): 208-224.

Posner, R.A. “Legal Formalism, Legal Realism, and the Interpretation of Statutes and the Constitution.” *Case Western Reserve Law Review* 37, no. 179 (1986).

Rachlinski, J.J., and Johnson, S.L. “Does Unconscious Racial Bias Affect Trial Judges.” *Notre Dame Law Review* 84, no. 3 (2009).

Rai, A. “Explainable AI: From Black Box to Glass Box.” *J. of the Acad. Mark. Sci.* 48 (2020): 137-141.

Re, R.M., and Solow-Niederman, A. “Developing Artificially Intelligent Justice.” *22 Stanford Technology Law Review* (2019).

Rosenthal, R. “Covert Communication in Laboratories, Classrooms, and the Truly Real World.” *Current Directions in Psychological Science* 12, no. 5 (2003): 151-154.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science* 300, no. 5626 (2003): 1755-1758.

Selbst, A.D., and Barocas, S. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87, no. 1085 (2018).

State v. Loomis, 881 N.W.2d 749, 760-761 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017).

"Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights," Council of Europe, last modified May 14, 2019. https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights.

Yu, S.B. "The Role of the Judge in the Common Law and Civil Law Systems: The Cases of the United States and European Countries." *International Area Studies Review* 2, no. 2 (1999): 35-46.

Završnik, A. "Criminal Justice, Artificial Intelligence Systems, and Human Rights." *ERA Forum* 20 (2020): 576-583.

Zerilli, J., Bhatt, U., and Weller, A. "How Transparency Modulates Trust in Artificial Intelligence." *Patterns*, 3, no. 4 (2022).

ORCID ID: Sofia Bonicalzi: 0000-0003-1335-2753

Abstract

In the last few years, decisions about matters of distributive and retributive justice have been more and more outsourced to automated systems (A.I.), and unprecedented ethical challenges have progressively emerged. As compared to human adjudicators, A.I.-based systems present, or may present in the future, concrete advantages in terms of efficiency and uniformity of performance. However, striving for uniformity may also have some sizeable costs. This paper aims to focus on a specific challenge – the difficult trade-off between uniformity and discretion in judicial applications of artificial intelligence – against the backdrop of current debates in philosophy, cognitive science, and artificial intelligence. I will argue that sidestepping the peculiarities of human reasoning might have some detrimental effects on the fairness of justice administration. This is particularly the case when the emphasis on uniformity is conducive to the elimination of reasonable standards of discretion, including the ability to bend the rules when circumstances so require.

Keywords: justice; opacity of algorithmic decision-making; cognitive biases; discretion in judicial decision-making.

Sofia Bonicalzi
Università Roma Tre
Cognition, Values, Behaviour Research Group,
Ludwig-Maximilians-Universität Munich
sofia.bonicalzi@uniroma3.it