Discrimination, Data-driven AI Systems and Practical Reason

Mireille Hildebrandt*

I. Distributive and Corrective Justice

In a constitutional democracy, governments owe their citizens equal respect and concern,¹ even though we cannot expect citizens themselves to feel equal respect and concern for each of their fellow citizens. The kind of equal respect and concern owed by governments to those subject to their rule is a form of *distributive justice*;² it is about similar cases being treated similarly to the extent of their similarity. This raises questions about which cases are similar and what factors are relevant to decide on similarity and thereby on difference. People are of course never entirely similar, so we need to know what difference *counts* and which difference should not be taken into account when governments decide on policies that affects their citizens and when they actually decide individual cases.

The kind of respect and concern citizens may feel for each other is a different matter, as we do not want to be forced into genuine concern for people we don't even know. Even those advocating a moral duty to engage with others based on equal respect and concern would agree that being more concerned about one's family or close friends than about others, is both understandable and legitimate for citizens, whereas it would amount to corruption if government officials were to treat their own family or friends differently from others when acting 'in function'. Citizens are, nevertheless, obliged to pay keen attention to *corrective justice* when interacting with others, respecting the mutual reciprocity that is at stake. We owe those who take care of us, who are dependent on our care, or those we trade with, work for, employ or hire appropriate respect and concern; not so much compared to others we engage with, but in relation to the relevant reciprocity – which differs between family members, between long term trading partners and between those who may never meet again.³ As should be clear, reciprocity between family members is based on relationships of dependence and trust; this is not about paying back immediately and precisely, and most often not about monetizable

DOI: 10.21552/edpl/2021/3/4

* Mireille Hildebrandt is a tenured Research Professor on 'Interfacing Law and Technology' at Vrije Universiteit Brussels. She works with the research group on Law Science Technology and Society studies (LSTS) at the Faculty of Law and Criminology. She also holds the parttime Chair of Smart Environments, Data Protection and the Rule of Law at the Science Faculty, the Institute for Computing and Information Sciences (iCIS) at Radboud University Nijmegen.

1 Ronald Dworkin, *Law's Empire* (Fontana 1991).

2 Aristotle, *Aristotle's Nicomachean Ethics*, Robert C. Bartlett and Susan D. Collins (trs) (University of Chicago Press 2012).

3 Marshall Sahlins, *Stone Age Economics* (Tavistock 1974).

exchanges. Reciprocity between, for instance, long term trading partners, employers and employees, doctors and patients, those who lend and those who borrow, is based on the need to safeguard long standing reliability; it involves myriad exchanges which are not necessarily commensurable even if money will often be used to 'make' them commensurable and thus resolvable. Reciprocity between strangers is based on the need to ensure precise and immediate compensation as the reciprocity is short lived; this usually implies agreement about monetization by way of a contract. It is crucial not to confuse distributive and corrective justice, and to distinguish the relationship between a government and its citizens from that between individuals and other private parties.⁴ Non-discrimination is about distributive justice and about organising society in a way that ensures equal respect and concern wherever distributive justice is at stake.

II. Direct and Indirect Discrimination

Direct discrimination refers to a *deliberate* attempt to exclude persons from what they deserve, desire or need, if such attempt is based on intended disrespect or a pronounced lack of concern for those persons, due to their political opinion, ethnic background, gender, religion or sexual orientation.⁵ Such deliberate discrimination may be at stake when equal cases are treated differently or when different cases are treated as if they count as the same. For instance, one could argue that if pandemic emergency relief is attributed to those who are self-employed based on their average trading profits over the preceding three year, direct discrimination is at stake where women who were pregnant during that period and therefor achieved a reduced average trading profit of X euros are treated in the same way as others who achieved X euros without being impacted by the pregnancy.⁶ Under the rule of law, governments are not allowed to engage in deliberate discrimination (a negative obligation) and to some extent they must see to it that private parties refrain from doing so (a positive obligation).

Indirect discrimination refers to treating people differently in a way that results in exclusion of persons due to their religion, political opinion, gender or ethnic background, even though there was no intent. In that case other factors may justify differential treatment. For instance, labour law may offer pregnancy leave to those who are pregnant (pregnancy is not a prohibited ground), thus excluding men from such leave (gender is a prohibited ground). This is obviously justified as long as men do not get pregnant. Based on this logic, parental leave would amount to direct – prohibited – discrimination if only available for women (because gender is a prohibited ground). Another example of indirect discrimination could be found in childcare if it were to be organised in such a way that persons of colour have substantially less access to it, for instance because of prohibitive cost, location or language issues. In effect, lack of access to

⁴ Jeremy Waldron, 'The Rule of International Law' (2006) 30 Harvard Journal of Law & Public Policy 15.

⁵ The prohibited grounds for discrimination differ between jurisdictions, but they are always closely related to liberty rights rather than notably economic rights.

⁶ See R. (on the application of Motherhood Plan) v HM Treasury [2021] EWHC 309 (Admin) (17 February 2021). This case is also highly relevant for indirect discrimination in times of emergency.

childcare reduces access to the labour market, which could mean disadvantaging or even excluding persons of colour from finding work. To make this point one would have to argue that persons of colour (a prohibited ground) more often have a low income (not a prohibited ground), more often live in neighbourhoods without proper childcare (not a prohibited ground) or more often cannot properly communicate with those taking care of their infants (not a prohibited ground), resulting in a substantially reduced chance to find work for persons of colour (prohibited ground). We can frame this by saying that low income, residence and availability of childcare in one's own language function as proxies for persons of colour. Depending on how they weigh in on access to the labour market for persons of colour, these proxies may indicate indirect discrimination. Those offering childcare may, however, justify their actions by pointing out they need to make a living too, that they are free to decide where to start their business and cannot be forced to only hire people who are fluent in the language of persons of colour. The pivotal question then becomes whether governments have a duty to prevent such indirect discrimination by reorganising the economic market for child care or by themselves offering childcare that is accessible in terms of money, distance and language for all their citizens.

Discrimination is also at stake when people are treated differently based on generalisations that may not hold for the individual but could nevertheless hold for a category they fit. Such unintentional discrimination may, for instance, target persons who live in a low-income neighbourhood as being a high risk for a life insurance. Statistics may demonstrate that – on average – those living in a low-income neighbourhood live a shorter life, thus increasing the risk for the insurance company. Corrective justice would allow the insurance company to charge a higher premium, as corrective justice plays out between two parties, not between a government and all of its citizens. There may be no intent to discriminate in the sense of showing disrespect or a lack of concern, but nevertheless those living a long life may end up paying a high premium if they live in a low-income neighbourhood. This kind of generalisation is core to intuitive human cognition and sometimes referred to as stereotyping.⁷ It is also core to any organic cognition (see below). In both cases the person or the organism that discriminates based on imperfect generalisation stands to be corrected; it may cost them their life if they apply the generalisation to the wrong target or – if human – they may be confronted with moral outrage from those wrongly targeted. Note that whereas we do not expect plants or elephants to give reasons for the way they discriminate, we do expect awareness and reflection from human persons. Institutions and AI systems apply this type of discrimination as a matter of course, even though such a policy may *indirectly* result in unintentionally excluding an individual person from what they deserve, desire or need. This is where the prohibition of discrimination becomes relevant. For instance, if – on average – persons of colour more often live in low-income neighbourhoods where insurance companies foresee a high risk, an individual person of colour who does not live in a low-income neighbourhood may nevertheless be charged a high premium. If being a person of colour

7 Frederick Schauer, *Profiles Probabilities and Stereotypes* (Belknap Press of Harvard University Press 2003).

is the only criterion this would amount to direct discrimination and be prohibited as such. If, however, the criterion is living in a low-income neighbourhood (economic status is not a prohibited ground) this could only amount to indirect discrimination if it results in disproportionately affecting persons of colour due to the fact that they more often live in low-income neighbourhoods. This could amount to indirect discrimination.

III. The End(s) of the Difference Between Direct and Indirect Discrimination

The introduction of data-driven AI systems to support or automate a range of decisions (from insurance to recruiting and from fraud detection to advertising) may disrupt the difference between direct and indirect discrimination. Data-driven AI systems are based on machine learning (ML), which develops mathematical functions that correlate input data with output data. For instance, data on residence, life-style, health records or educational background may correlate in different ways with claim-behaviour, thus offering an assessment of the risk for an insurance company. This assessment will influence decisions on acceptance and on price. The point is that ML systems necessarily work with datified *proxies* (variables), because computational systems cannot work with concepts like ‘health’ or ‘education’. They require formalisation in the form of *discrete variables*, for instance ‘health’ can be replaced by a preformatted listing of health conditions. If such a variable correlates with a prohibited ground of discrimination, this may be an indication of bias in the sense of indirect discrimination. The problem is that this bias can only be established if the correlation with prohibited grounds is known, whereas the collection of sensitive data needed to check for such bias may not be available (collecting such data may even be prohibited to prevent direct discrimination).⁸

Some ML systems, known as ‘deep learning’ (DL), develop highly sophisticated, dynamic and complex algorithms to map the mathematical relationships between a massive number of variables and the targeted output. Here, the fluidity and complexity of the algorithm raises the question of which variable is a proxy for what other variable, and whether that relationship is a one-way street, as the distinction between direct and indirect discrimination assumes. Direct discrimination means that a variable that stands for a prohibited ground is used to discriminate, whereas indirect discrimination may be at stake if another variable is used that, however, correlates with the variable that stands for a prohibited ground. For instance, discrimination based on a variable such as gender is prohibited as direct discrimination. If, however, discrimination is based on a proxy variable such as a type of job that is mainly held by either women or men, we are in the realm of indirect discrimination. What happens if DL systems enable developers to play around with variables in a way that obfuscates the relationship between prohibited grounds on the one hand and highly complex combinations of variables on the other hand, noting that the prohibited ground may not even be listed as a variable? My sense is that the use of DL systems will make the distinction between direct and indirect discrimination an illusion, with users of these algorithms claiming

⁸ Art. 10.5 of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act) and amending certain legislative Acts, 21 April 2021 COM/2021/206 final.

that their decisions are justified by high-accuracy and blind to any kind of bias (because the prohibited variable is not in the training data). Though the latter obfuscates bias rather than precluding it, this argument is often used to claim neutrality, confusing the prohibition of direct discrimination with that of indirect discrimination. High time for lawyers and legislatures to rethink the relevance of the difference when data-driven AI systems have been deployed, taking note that the same does not apply to code-driven AI systems based on symbolic logic.⁹

IV. Economic Deprivation is Not a Prohibited Ground

The prohibition of direct or indirect discrimination addresses unequal treatment, but only if based on a limited set of factors, such as religion, ethnicity, skin colour, trade union membership, criminal conviction, sexual orientation or gender. Deploying these factors allegedly violates human dignity, highlighted by its central place in the Universal Declaration of Human Rights and subsequent legal instruments (both constitutional and international). Though the concept of human dignity defies definition it should at least be understood in terms of the intricately interwoven concerns of liberty and equality (instead of playing them against each other); a person who is treated unequally in a way that effectively excludes them from access to notably education, employment, credit or healthcare, will be less free due to suffering the consequences of unequal treatment. Not only because their options have been reduced (the reductive utilitarian approach) but also because they cannot develop their capabilities the way others can; their ability to navigate their world is more restricted because they have not been free to initiate or improve such abilities (Sen's capability approach).¹⁰

In market economies discrimination plays a key role; providers of products and services aim to discriminate between audiences that may and audiences that may not be willing to buy or use their products and services, between what different categories of consumers are willing to pay for their products, and between more and less effective ways to reach and persuade different types of *potential* consumers. On top of that they need to discriminate between their own offers and those of their competitors, while anticipating the effects of price discrimination and quality standards on their market share (inviting trade-offs with potentially perverse implications). Innovation, therefore, aims to stay one step ahead of the competition, claiming novelty to lure consumers away from competitors (enlarging the market and/or enlarging the market share). Advertising and marketing play a major role in the creation, capture and transformation of markets and market share. The key players here are advertisers (those offering products and services), publishers (newspapers, television, social media, search engines and other online platforms who get paid to host advertisements), readers or visitors (who are confronted with paid content that is aimed to influence them) and advertiser intermediaries. The role of the latter has become decisive for much of the paid content and structure of online

⁹ They nevertheless raise other issues, see the case law referred in note 6 above. For an in-depth exploration of these issues Zenon Bankowski, Ian White and Ulrike Hahn (eds), *Informatics and the Foundations of Legal Reasoning* (Springer 2013).

¹⁰ Amartya Sen, 'Elements of a Theory of Human Rights' (2004) 32 *Philosophy & Public Affairs* 315-56.

publications that obtain a fee for hosting advertising. Influencing has become equivalent with AB testing and behavioural profiling, grounded in cross-contextual tracking and tracing. This also goes for recommender systems and search engines that use similar technologies to frame online content in ways meant to prioritise lucrative behaviour, taking note of the fact that the same eco-system of nudging combined with machine learning has been instrumentalised to influence political opinion, voting – if necessary, based on deep fakes and other types of dedicated misinformation.

Price discrimination is not prohibited; neoliberal ideology tells us that - if certain conditions apply - price discrimination will contribute to efficiency in economic markets.¹¹ It may, however, result in direct or indirect discrimination if based on prohibited grounds, or lead to exclusion or detrimental treatment for those who are part of a protected category (higher prices will at some point lead to exclusion). Price discrimination may concern the price of labour (a salary), of insurance (the premium), of education (the fees) or of healthcare (again the premium). This may in turn result in direct or indirect discrimination if – for instance – women are systematically offered a lower salary, or if persons of colour find themselves qualified as high risk for various insurance products because of the fact that many people of colour live in low-income neighbourhoods that are redlined as high risk, resulting in them having to pay a high premium. Even if they are not low income themselves and do not live in a low-income neighbourhood, based on the statistics they may be ‘branded’ as such and suffer the consequences. If this was intended it may be a form of prohibited direct discrimination – if not, it may still be indirect discrimination that, insofar as not justified, would be prohibited.

This also means that low-income folk may generally pay higher premiums, because low economic status is not usually a prohibited ground. So those with less money pay more, thus reinforcing existing inequality and contributing to less freedom for those already disadvantaged. Much discrimination is not prohibited, because it is based on factors such as economic deprivation rather than political opinion, ethnicity, sexual orientation or gender. Unless low economic status (1) disproportionately affects one of the prohibited grounds, (2) can be proven to do so and (3) cannot be justified, discrimination based on low economic status is not prohibited. Though this may satisfy corrective justice between insurer and insured (if it concerns a higher premium in case of a higher risk), and even satisfy distributed justice between an insurer and its clients (if all clients with the same risk pay the same premium), it raises questions from the perspective of distributive justice in the overall context of a society. I will return to this at the end of this foreword.

V. From Organic to Human to AI Profiling

Gregory Bateson, one of the founding fathers of cybernetics (and of what is called AI these days), discussed information in terms of ‘the difference that makes a difference’,

11 Excellent introductions to the provenance of neoliberal economic thought, highlighting the activist and ideological background of those advocating and inducing neoliberal policies: Philip Mirowski and Dieter Plehwe (eds), *The Road from Mont Pelerin: The Making of the Neoliberal Thought Collective* (1st edition, Harvard University Press 2009); Lina Khan, ‘The Ideological Roots of America’s Market Power Problem’ (2018) 127 Yale Law Journal, 960-979.

or ‘negative entropy’.¹² Bateson noted that when we perceive our environment we can discern an infinite number of differences between things, and to survive and flourish we need to discern the *differences that make a difference* (to us). If we get this wrong, as a person, as a community, as humanity, we may not survive and will not flourish. In fact, all living beings are in the process of discerning the difference that makes a difference in their environment, as this is what affords organisms to act on relevant differences. Organisms thus continuously anticipate what differences will impact them and how they can respond in a way that sustains and nourishes their life. As I wrote in *Profiling the European Citizen*,¹³ learning to discriminate between what is relevant and what is not relevant is a critical sign of life.

My point was that if ‘AI systems’ try to anticipate and even pre-empt our behaviours, they are doing something that has been core to ‘organic profiling’, human profiling and institutional decision-making. Discriminating what matters from what is irrelevant has been key to individual organisms, to types of organisms, to individual persons and communities, and to human society as a whole. Basically, discrimination will often be based on intuitive generalisation, and whereas ‘getting things right’ when generalising is a matter of survival and flourishing for all living organisms, much more is at stake at the level of human society. Human society can survive and maybe even flourish based on very different structures that imply very different types of discrimination, including those that affront human dignity by restricting the freedom of some compared to broad leeway for others.¹⁴ If constitutional democracies take human dignity seriously, the discrimination that is key to human survival and flourishing must be done in a way that demonstrates equal respect and concern. This necessitates a concomitant discernment and acuity to be built into AI systems wherever they make or support decisions that make a relevant difference. Following Bateson, we need to discern *the difference that makes a difference* between the way that AI systems discriminate and the way that organisms and human life forms discriminate, to answer the question whether AI systems are capable of distinguishing between behaviour that shows equal respect and concern and behaviour that ignores this kind of distributive justice. If we frame distributive justice as fairness, we can frame this question as one of fairness by design: can we design automated decision making (ADM) and automated decision support (ADS) systems such that they avoid both direct and indirect exclusion from what people deserve, desire or need?

VI. Discrimination and Discernment: Practical Reason, Logic and Experience

Equal treatment is easier said than done. It is not an engineering problem, though ADM and ADS systems have now made it so. To treat people equally and to have people

12 Gregory Bateson, *Steps to an Ecology of Mind* (Ballantine 1972), 453.

13 Mireille Hildebrandt, ‘Defining Profiling: A New Type of Knowledge’ in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen. A Cross-disciplinary Perspective* (Springer 2008), 17-45.

14 Katharina Pistor, *The Code of Capital: How the Law Creates Wealth and Inequality* (Princeton University Press 2019). Ian Law, *Red Racisms: Racism in Communist and Post-Communist Contexts* (Palgrave Macmillan 2012).

treated equally, we need to decide on *what counts as being equal*, which means that we need to agree what factors/grounds/variables are relevant in what context, based on what reasons. In other words, to prevent prohibited discrimination we need to *discriminate* between myriad factors/dimensions/affordances of an event or a situation – in a way that allows us to *anticipate* what discrimination is lawful. This means that distributive justice is not only about treating a case similar to previous cases, but also to future cases. This is what enables people to plan their life: creating legitimate, reasonable expectations.

To discriminate lawfully requires *discernment*, a concept that refers to *acuity* in perceiving relevance in the face of both past experience and future expectations. Discernment relates to what Bateson called the ‘knitted’ texture of knowledge, which is ‘all sort of knitted together, or woven, like cloth, and each piece of knowledge is only meaningful or useful because of all the other pieces’.¹⁵ Discernment and acuity also relate to what Aristotle and virtue ethics call ‘practical reason’ or *phronesis*, the trained intuition of skilled practitioners capable of judgement rather than mere proficiency in logic. This also distinguishes discernment from calculation and computation; it is not just about the ability to sum up hundreds, thousands or billions of variables that may correlate with an outcome we define as ‘fair’, ‘just’ or ‘lawful’, but about the capability of teasing out the most pertinent dimensions of an issue that cannot be reduced to a preconceived and formalised outcome. ML algorithms cannot be trained on future data, ML research design must assume that the distribution of future data is similar to that of historical or streaming data. As Mitchell has clarified,¹⁶ this assumption does not fly – even though it is helpful from an engineering perspective. Practical reason enables us to imagine the future in terms of the past, and the past in terms of the future – a reflective equilibrium well known to any practicing lawyer when deciding a case. It entails discretion, which should however not be used in an arbitrary way. As Dworkin argued, discretion is the heart of the law.¹⁷ It requires constructive interpretation tuned to the integrity of the law and it is precisely this particular type of construction that is beyond the logic of ML systems. In other work I have framed the issue of ADM and ADS in law as ‘freezing the future and scaling the past’;¹⁸ fairness is not computable, it requires keen awareness of potentially relevant proxies for prohibited grounds of discrimination and creative intervention to link them to past judgments.

Does this mean that we should reject the entire domain of ‘fair computing’? No, on the contrary, I would say we have so much to learn. Facing the formalisation of different conceptions of fairness in the context of data-driven ML systems offers a mirror of how fairness can be or has been decided and may thus help us to reflect on how we did or may frame the difference that makes a difference. If we use AI systems to face

15 Bateson (n 12), 21-22.

16 Thomas Mitchell, *Machine Learning* (McGraw-Hill Education 1997), 6.

17 Ronald Dworkin, *Taking Rights Seriously* (Harvard University Press 1978).

18 Mireille Hildebrandt, ‘Code-Driven Law: Freezing the Future and Scaling the Past’ in Christopher Markou and Simon Deakin (eds), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing 2020).

that mirror instead of stepping into an echo chamber or filter bubble we may improve our capability to move away from discrimination.¹⁹ This means that data-driven ADM and ADS systems should be reconfigured to *not decide or support* our decision making but *to confront us* with past and potential future discrimination, thus enriching the kind of self-reflection that is key to non-discrimination.

19 Mireille Hildebrandt, *Algometrisch strafrecht: Spiegel of echoput?, Delikt en Delinkwent* (forthcoming September 2021).