# Algorithms and Values in Justice and Security

Hayes, Paul; van de Poel, Ibo; Steen, Marc

**DOI**
10.1007/s00146-019-00932-9

**Publication date**
2020

**Document Version**
Final published version

**Published in**
AI&Society: the journal of human-centered systems and machine intelligence

**Citation (APA)**
Hayes, P., van de Poel, I., & Steen, M. (2020). Algorithms and Values in Justice and Security. *AI&Society: the journal of human-centered systems and machine intelligence*, *35*(3), 533-555. https://doi.org/10.1007/s00146-019-00932-9

AI & SOCIETY
https://doi.org/10.1007/s00146-019-00932-9

ORIGINAL ARTICLE

# Algorithms and values in justice and security

Paul Hayes[1] · Ibo van de Poel[2] · Marc Steen[3]

Received: 2 April 2019 / Accepted: 10 December 2019

**Abstract**

This article presents a conceptual investigation into the value impacts and relations of algorithms in the domain of justice and security. As a conceptual investigation, it represents one step in a value sensitive design based methodology (not incorporated here are empirical and technical investigations). Here, we explicate and analyse the expression of values of accuracy, privacy, fairness and equality, property and ownership, and accountability and transparency in this context. We find that values are sensitive to disvalue if algorithms are designed, implemented or deployed inappropriately or without sufficient consideration for their value impacts, potentially resulting in problems including discrimination and constrained autonomy. Furthermore, we outline a framework of conceptual relations of values indicated by our analysis, and potential value tensions in their implementation and deployment with a view towards supporting future research, and supporting the value sensitive design of algorithms in justice and security.

**Keywords** Values · Value sensitive design · Responsibility · Ethics · Algorithms · Justice · Security · AI

## 1 Introduction

Algorithms are powerful artefacts that operate within our informational milieu, structuring our data, profiling, categorizing, and predicting who we are, what we want and more. These artefacts are becoming increasingly authoritative for the insights they produce, and the promises they bear for decision support and resource management. The governance model we are drifting towards has been argued to variously be an "algocracy" and before that infocracy, which is perhaps emblematic of the potential for the diminishing role and autonomy of the human decision-maker as information production and decision-making become increasingly automated, authoritative, and opaque (van den Hoven 1998; Danaher 2016, 246–248; Peeters and Schuilenburg 2018). Here, we are concerned with the uses of algorithms in the area of justice and security, a particularly sensitive context with great potential to benefit from their power to produce insights to help enforce the law, but also a significant capacity to cause harm.

Generally in the area of policymaking, Big Data (and by extension algorithms) "…can support evidence-based policymaking" and "…can help officials make better decisions and improve government efficiency and effectiveness" (van der Voort et al. 2019, 27). Algorithms in the domain of justice and security can serve many purposes such as identifying people at risk of gun violence (as subject or perpetrator), identifying geographical areas at heightened risk of crimes including burglaries, license plate and facial recognition, likelihood of recidivism, child welfare and safety, and many more (Police (UK), nd; Angwin et al. 2016; Garvie et al. 2016; O'Neil 2016; Ferguson 2017b; Eubanks 2018). What's

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00146-019-00932-9) contains supplementary material, which is available to authorized users.

✉ Paul Hayes
p.d.hayes@tudelft.nl

Ibo van de Poel
I.R.vandepoel@tudelft.nl

Marc Steen
marc.steen@tno.nl

[1] Ethics and Philosophy of Technology, Values Technology and Innovation, Faculty of Technology, Policy and Management, TU Delft, Building 31, Room number: B4.060, Jaffalaan 5, 2628BX Delft, The Netherlands

[2] Ethics and Philosophy of Technology, Values Technology and Innovation, Faculty of Technology, Policy and Management, TU Delft, Building 31, Room number: B4.210, Jaffalaan 5, 2628BX Delft, The Netherlands

[3] Human Behaviour and Organisational Innovations, TNO, New Babylon, Anna van Buerenplein 1, 2595DA Den Haag, The Netherlands

Published online: 11 January 2020

more, the gaze of authority and smart number crunching need not simply be turned on the governed population. In at least the policing context, the data that police officers generate can potentially also be processed and modelled to create insights on how to improve police performance and accountability (Ferguson 2017b, 143). Ferguson (2017b, 143) calls this accountability driven data use "Blue Data."

Algorithms and AI have the potential to create opportunities for human dignity and flourishing (Floridi et al. 2018), however there also exists the potential for misuse, and more pragmatically underuse stemming from "…fear, ignorance, misplaced concerns or excessive reaction…" to misuse or abuse (Floridi et al. 2018, 691). For every area of legitimate opportunity that algorithms and AI provide, there are countervailing risks of harm (Floridi et al. 2018).[1] We adhere to a balanced view. In our broader research, we are exploring which values are at play in the design, implementation and deployment of algorithms with the aim of understanding how to maximise their contribution to human dignity and flourishing whilst minimising their potential for misuse. In order to do this, we are attempting to discern how algorithms can uphold our moral values by investigating their value impacts, and how corresponding norms can be translated into their design. This approach is known as value sensitive design (VSD)—a tripartite methodology involving conceptual, empirical, and technical investigations of a studied technology (Friedman et al. 2013).

Values are associated with what is good or (objectively) desirable (van de Poel 2018, draft; Schwartz and Bilsky 1987; Friedman et al. 2013). They are evaluative (van de Poel 2018, draft) and help us to evaluate current states-of-affairs against those that are ideal. They are often not directly action-guiding but they may be associated with norms and ultimately with design requirements for technical and institutional systems, so that they can help in the design and use of value sensitive algorithms (van de Poel 2013). In this article, we aim to identify the main values that are relevant for algorithms in justice and security and how they support each other or conversely, come into tension.

The identification of key values is of unique importance in justice and security, where misuse of an algorithm could undermine values and come at a large cost to our freedoms. Fear of this misuse may also cause underuse (Floridi et al. 2018, 691)—public scepticism could unreasonably hinder the development and implementation of artefacts that have the potential to uphold the values of justice, security, human flourishing and welfare more generally by providing invaluable, potentially life-saving, assistance in decision-support to agents of the state in enforcing the law and providing security. The answer to the problem of preventing misuse, and underuse, will be intentional design that is cognisant of our human values.

In what follows, we identify key values and unpack their relevance and implications for algorithms in justice and security. We take and analyse broadly the theoretical and documented implications of algorithms on seven values; accuracy, autonomy, privacy, fairness/equality, ownership/property, and accountability and transparency. We argue under each heading that there are significant risks arising to these values, or in some cases potentially from them as they interact with others (for instance, in practice ownership can be deleterious for transparency).

In order to help designers in particular mitigate and weigh these risks accordingly, we propose a framework of the conceptual support of values and their tensions in implementation. Such a framework can support reflection on values during the design process.

## 2 Value implications of algorithms for justice and security

In what follows, we will discuss several values that are relevant for the design, implementation and deployment of algorithms in justice and security for a number of key values, as above stated: accuracy, autonomy, privacy, fairness/equality, ownership/property and, accountability and transparency. Before we discuss these values in more detail, a few words need to be said about why we have selected this particular set of values.

Our focus is on values that are morally important for the design of algorithms that are used in the domain of justice and security. This means that the relevant values are both determined by the object of design (i.e. algorithms) as well as the domain of application (i.e. justice and security). Moreover, we are interested in moral values, or at least values of moral importance.

Concerning algorithms, we take inspiration from the four ethical principles that High-Level Expert Group of the EU on AI has formulated: respect for human autonomy, prevention of harm, fairness, and explicability (High-Level Expert Group on AI 2019). More generally, we have looked at values that have been identified (in VSD) as being relevant for the design of information systems: human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, courtesy, identity, calmness, and environmental sustainability (Friedman et al. 2006).

[1] Opportunities, according to (Floridi et al. 2018, 691) include "Enabling human self-realisation", "Enhancing human agency", "Increasing societal capabilities", "Cultivating societal cohesion", whereas corresponding risks include "Devaluing human skills", "Removing human responsibility", "Reducing human control", and "Eroding human self-determination."

In delineating the set of relevant values, we also took into account the specificities of the domain of justice and security. The underlying idea here is that in different societal domains different values are of prime importance. Philosophers have formulated this idea in terms of different societal spheres of justice (Walzer 1983) or of different societal institutions being aimed at the realisation of different kinds of ends or of collective goods (Miller 2009). For example, with respect to the police, Seumas Miller (2009, 245–246) states that the "central and most important purpose, that is, collective end, of police work is the protection of moral rights, albeit this end, and its pursuit by police, ought to be constrained by the law." What is important about this formulation is that police work is not only about safeguarding the moral rights of (potential) victims, but—by its nature—also requires respecting the moral rights of (potential) suspects and perpetrators. What Miller says here about the police seems to apply more generally to institutions in the domain of justice and security. Under the rule of law, this has been translated into such legal rights as the presumption of innocence, the right to due process, and the equal and fair treatment of people. In our value framework, these considerations are translated into the values of autonomy, fairness (and equality) and privacy.

Other values in our framework derive from the fact that we focus on algorithms. One of the main ethical concerns with respect to algorithms is their potential opacity (Mittelstadt et al. 2016). Such opacity may result in three types of moral problems. First, it may result in decisions being made that lack explainability and, hence, lack a clear justification. While the ability to justify (important) decisions is important in general, it is crucial in the domain of justice and security. In our framework, this translates into the values of transparency, and accountability. Second, opacity may result in a lack of responsibility (and accountability) for the decisions being made. Again, this is particularly important in the domain of justice and security. Here the value of accountability is important but also values like the autonomy of the decision-makers, and ownership and property, as ownership has implications for legal and moral responsibility (Robaey 2015). Third, opacity may also result in bad or wrongful decisions; here particularly the value of accuracy is relevant.

Our claim then is that our list of values is particularly important if one wants to properly address the moral concerns that the use of algorithms in the domain of justice and security raise. This does not mean that we claim that our list is exhaustive. We also recognize that the values may be named and grouped differently, but we believe that our current presentation most clearly foregrounds values of moral importance.

In addition, we would like to point out that the values that we discuss (below) are likely to be interpreted differently when viewed through the lenses of different ethical traditions. A discussion of such different interpretations is outside the scope of our paper. Several examples for one value, however, may illustrate what we mean. Through a consequentialist lens one might evaluate a value like accuracy in terms of the consequences that follow from using the algorithm's outcomes, for example, the (probably positive) consequences of an algorithm's true positive and true negative outcomes, for example, the increase in public safety of correctly detecting criminal behaviour. Or the (probably negative) consequences of an algorithm's false positive and false negative outcomes, for example, the costs of inefficiency of having to correct and repair these errors. Through a deontological lens one might evaluate accuracy differently, for example, in terms of a police organization's duties to protect citizens against criminal behaviour, or in terms of upholding human dignity of citizens, a duty to treat each person as innocent until sufficient proof of guilt is gathered and tested. And lastly, a proponent of virtue ethics would evaluate accuracy by looking at the ways in which usage of an algorithm enables or hinders police officers to do their work properly. Is the algorithm's accuracy good enough to support police officers in cultivating virtues like honesty and trustworthiness, in how they approach and treat citizens? Or is the accuracy so poor that police officers feel that using the algorithm would corrode their honesty and trustworthiness? A virtue ethicist may also zoom-out to the level of society and ask whether this algorithm, given its accuracy, helps or hinders to create a society in which people can flourish.

In our discussion of the values (below), we will remain agnostic of these different ethical lenses and interpretations and, effectively, follow a pluralist approach.

### 2.1 Accuracy

Accuracy, in our frame of analysis, can thinly be understood as fidelity or closeness to truth. In practice, in big data and data analytics this can be operationalised more thickly. In the data quality literature, it is also associated with completeness of data, consistency of format, relevance, and timeliness (Fox et al. 1994; Tayi and Ballou 1998, 56). This is relevant to our discussion here, as algorithms not only use data for their insights, but produce it also. Other dimensions of data quality relevant to accuracy include precision (the measurement standard) and reliability (or probability of correctness) (see Fox et al. 1994, 14–15). In more complex terms, accuracy has been defined by Christopher Fox et al. (1994, 14) as:

> …the degree of closeness of its value *v* to some value *v'* in the attribute domain considered correct for the entity *e* to the attribute *a*… If the datum's value *v* is the same as the correct value *v'*, the datum is said to be accurate or correct.

Should an algorithm's inputs or training data be using inappropriate measurements, be incomplete, or be unreliable, or poorly maintained or not sufficiently purged of error (cleaned), these failures in data quality will likely lead to unacceptable error rates. Furthermore, we want the output data of our algorithms to be relevant and timely, as well as being an appropriate measure for the phenomena they are to provide some insight on. Accuracy then is a property of training data, input data, and output data.

Additionally, an algorithm's accuracy can also be threatened by poor data entry practices, policing practices, outlier events, and model overfitting (see McCue 2015, 17–18, for more on the latter two). Human choices made in design (and even choices made before design) will have an impact on data quality and an algorithm's accuracy. We will explore this more broadly in the following.

We want our data to reflect reality, particularly when we are basing important decisions on it—in this context decisions including where to send police patrols, who to target for police intervention, who spends how long in prison, and more. If our information is not appropriately accurate for its goal it is not very useful and is potentially dangerous. Supplied with inaccurate information, we may make inappropriate, ineffective, or harmful decisions. When an algorithm produces falsities, it provides red herrings, not actionable insights. It becomes the antithesis of its creators' and users' presumed good intentions. However, we say appropriately accurate as the value of accuracy is complex in practice. Data scientist Coleen McCue (2015) points to some of the nuances of accuracy.

Firstly, McCue (2015, 8) points out that a very high degree of accuracy may not be very useful for low frequency events—McCue (2015) gives the hypothetical example of an algorithm that predicts the escalation of robbery into assault with only one decision rule, "no." Such an algorithm could feasibly be correct 95% of the time but of course would be useless (McCue 2015, 8). Secondly, McCue (2015, 18–19, 268) argues that there is a trade-off between accuracy and, generalizability and interpretability, that is, in some contexts highly specific or complex information may not be actionable. In an example given by McCue (2015, 18–19), a hypothetical predictive algorithm that can allocate risk scores for geographic areas in 30 min blocks might be highly accurate, but very challenging to act upon.

Whilst technical artefacts such as algorithms might be regarded as objective and impartial, even infallible for the technologically naive, they are only as good as the data on which they are trained. The widely held consensus is that algorithms are oftentimes not impartial or objective, and are imbued with human biases (to be explored below) or goals and ideology, either as a design decision or due (whether conscious or subconscious) to overrepresented or underrepresented data subjects, or erroneous data (O'Neil 2016; Ferguson 2017b; Kitchin 2017, 17–18). Cathy O'Neil (2016, 20–21) describes algorithmic models as simplifications that cannot capture all of the world's complex phenomena, and when constructing models choices necessarily must be made about what data to include in these simplified models of the world (or part of it), which leads to blind spots. Kelleher and Tierney (2018, 47–48) emphasise the crucial importance of attention to the design of data abstractions, data quality, and a critical approach to results of the process as identified patterns may not be real insights, but reflections of "…biases in data design and capture."

Ferguson (2017b, 52) indicates that such issues may be compounded in the justice and security setting (specifically the area of policing), arguing that "…in fact, because of the volume of data coming in, the complexity and the lack of resources to cleanse and correct mistakes, these systems are more likely to contain mistakes." Additionally, some databases were simply not designed to be analysed (McCue 2015, 82) and the data therein may not easily serve statistical analysis.

Furthermore, with certain types of crime underreported (in some cases, potentially due to low trust or confidence in the police[2]) or misreported, the source data underlying algorithmic models may preclude accurate analysis (Ferguson 2017b, 72; Richardson et al. 2019, 201). Moses and Chan (2018, 809) add that data may not always be categorised consistently or accurately, and as predictive policing in particular will influence data collection itself, "[t]his feedback loop is self-perpetuating, potentially resulting in observed stability of crimes, locations and individuals monitored by police despite potential changes in the actual crimes committed." O'Neil (2016, 87) calls this a "pernicious feedback loop" whereby the resultant focused policing creates new data, which then ostensibly justifies the policing pattern observed. This is not a new phenomenon, being similar to a "ratchet effect" (Harcourt 2005, 27), but will be an increasingly pertinent risk with continued reliance on algorithmic and actuarial practices.

There is empirical evidence supporting the claim of the pernicious feedback loop as it applies to algorithms in justice and security. Using PredPol's geo-spatial risk based algorithm (it being one of few publicly released in a peer-reviewed journal), Lum and Isaac (2016) tested National Survey on Drug Use and Heath Data (NSUDH) against police arrest data relating to drug use in Oakland, California. The different data sets, as visualised on maps, told different stories, with arrests concentrating in non-white low income areas, whilst NSUDH data showed a more even distribution of drug use (Lum and Isaac 2016, 17). Using the PredPol

[2] This is, however, a complicated phenomenon that has produced some varying evidence (Kääriäinen and Sirén 2011; Boateng 2018).

predictive algorithm, Lum and Isaac (2016, 18) found that these areas overrepresented on the police database continued to be overrepresented in drug related crime predictions.

Richardson et al. (2019) refer to the phenomenon of inaccurate or biased data collection as "Dirty Data", data which may reflect poor policing practices including racial discrimination, misreporting and other misconduct, and argue that such data can also arise as a result of corrupt practices, thus exacerbating the so-called pernicious feedback loop.

Current research on the state-of-the-art of the accuracy of some algorithms used in justice and security is mixed. Take an interesting example from Chicago. Based on analysis of homicide statistics gathered between March 2013 and March 2014, Saunders et al. (2016, 362) found (in relation to the Strategic Subjects List, or SSL[3]) that:

> …0.7% of the SSL subjects were homicide victims, 0.4% of the 17,754 associates were homicide victims, 0.029% of the 855,527 former arrestees with no associates were homicide victims, and 0.003% of the rest of the almost 2 million Chicago residents without any criminal record were victims of homicide.

These statistics suggest a rather limited capture of victims of gun violence, however, Saunders et al. (2016, 366) emphasise that persons on the SSL were nonetheless 233 times more likely to be homicide victims than the average Chicago resident. On the one hand, this algorithm's predictions would seem to have resulted in few actuations of events based on risk, however by comparison to the average population the risk calculations would appear to be well justified. Additionally, Saunders et al. (2016, 366) report that later statistics compiled by Lewin and Wernick (2015) (a member of the Chicago Police Department and the algorithm's designer) show that "…29% of the top 400 subjects were accurately predicted to be involved in gun violence over an 18-month window." Subsequent increases in gun violence in Chicago resulted in criticism of the algorithm (Ferguson 2017b, 39).

A key question arising from something such as the SSL becomes whether it is acceptable to place these people on such a list if there is a possibility of it increasing negative encounters with the police.

Algorithms (and technical artefacts more broadly) are argued to be performative within their socio-technical assemblages, influencing agents to take action based on their outputs as a sometimes unquestioned authority (Niculescu Dinca 2016; Kitchin 2017, 19). If the data and output are bad, then it should come as little surprise if a performed action it inspires is ineffective or harmful, persons falling victim to false negatives (or living within an neighbourhood flagged as high-risk) will likely have deleterious contacts and experiences with the justice and security system.

Though the system resulted in no known wrongful deaths or reported incidents, Ferguson (2017b, 84) describes how Fresno California Police piloted a programme called Beware, which "…searches through proprietary consumer databanks to provide a rough predictive judgement [colour coded threat levels] about a 911 caller, the address, or the neighbourhood." During a public hearing about the system, a local councilman asked for his address to be run through the system, only to find that his house was considered a non-trivial yellow threat (Ferguson 2017b, 85).[4] Ferguson (2017b, 85) explains that whilst the man was not a known threat, in responding to a call police officers (presuming they had no additional information) would likely have regarded him with caution. This anecdote offers a useful insight into how reality might be distant from the data which themselves purport to be insights that may influence the interactions of the police and the algorithm's subjects. Take for example another cautionary tale further reflecting these risks. Ferguson (2017b, 95) describes how a licence plate misread by an automatic number-plate recognition (ANPR) technology resulted in a 47-year-old African–American woman being stopped by the police at gun point. This anecdote provides a rather firm example of the shape of potential dangers of civilian and police interactions that are mediated by artefacts, how inaccuracy can instigate unwanted and unjustified police contacts.

Indeed, the use of image processing algorithms deserves special mention when so many major cities today are equipped with CCTV cameras incorporated into algorithmically empowered systems for image recognition (ANPR, facial, and gait recognition) (Kitchin 2016, 7). Introna and Wood (2004, 188, 190–191), in a reasonably through analysis of the politics of facial recognition algorithms outline their historical and significant vulnerabilities, noting algorithmic performance can degrade depending on the size of databases and the age of photographs used for matching (demonstrating the importance of the timeliness of data used), as well as disparities in matching rates or recognisability by race and gender.

The value of accuracy, as it relates to truth, is good in itself. However at the point that actions are performed based on falsities, other values are implicated. Fairness or equality

[3] The SSL is a predictive algorithm which "…uses 11 variables to create risk scores from 1 to 500" where an individual is more likely to be a victim or perpetrator of gun violence the higher their score (Ferguson 2017b, 37).

[4] In another case of data error, Ferguson (2017b, 49) reports that a California auditor found a police database that identified 42 infants as gang members.

is a major value where biased data is used to train an algorithm, and this will be explored below.

When simplified models are made about the world, choices have to be made about what data to include, and what data to exclude (choices that often remain implicit and therefore unexamined), and in other cases data that might be useful is either not available or cannot be formalized in a manner understandable to a computer—nuance is lost and variables that might otherwise alter an algorithm's decision are not analysed (Angwin et al. 2016; O'Neil 2016, 20–21; Eubanks 2018, 147).

The point of the preceding is not to condemn algorithms, but outline the risks. An accurate algorithm can be a useful resource in informing effective decision-making. Nevertheless, we need to hold a serious discussion about what kind of threshold of accuracy is acceptable when these algorithms can have recursive and potentially powerful impacts on a host of our values. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)[5] algorithm for instance, according to an in-depth investigation using data from Broward County, Florida, ProPublica found the algorithm to have a 61% rate of correctly identifying recidivists (Angwin et al. 2016). This makes it only just a more accurate tool than no tool at all but with potentially serious consequences for victims of false negatives (slightly above 50% rate of the proverbial coin flip), leaving 39% of offenders apparently unnecessarily identified as at risk of re-offence (Angwin et al. 2016). This algorithm makes decisions with weighty consequences despite its accuracy being in question.

In addition to this, recent research by Dressel and Farid (2018, 2–3) found that a small group of non-experts could predict recidivism at a similar level of accuracy as COMPAS with less information (seven versus 137 features) and furthermore the authors found that "[a] classifier based on only two features—age and total number of previous convictions—performs as well as COMPAS."[6] This research raises a concern about the wisdom of implementing data hungry[7] algorithms that may not add value to the processes for which they were designed, and underscores that we must carefully evaluate and reach some consensus on an acceptable threshold of accuracy before implementation, that is at least fitting to the particular context of use. Some make the reasonable argument that algorithms such as COMPAS should not be deployed in their particular contexts, but rather utilised in alternative ones such as prioritizing persons in need of "…more services and support in the re-entry process" (D'Ignazio and Klein 2018, draft). This is an important point given their potential to cause harm to offenders with little apparent value added to the decision process.

In contrast, some successes demonstrate that, at least combined with effective implementation or incorporation into tactical and strategic planning, algorithmic deployment can translate into positive results (see Perry et al. 2013). Perry et al. (2013, 43–44) for instance report decreases in property crimes by 19–20% in two California police districts during observed time periods where algorithms were used.

Ultimately, on top of ensuring the quality of training and input data, those working with data and designing algorithms must use methods to determine what kind of errors can be expected (such as confusion matrixes), whether the risk of false positives is appropriate to the context of use, and relatedly whether accuracy can be compromised for generalizability and interpretability (McCue 2015, 9, 160). Additionally, considerations need to be made to eliminate bias in design that could skew results unfavourably against particular groups or individuals (a topic to which we will soon return).

In sum, algorithms are only as accurate as the data they are trained on in conjunction with design decisions made along the way. Training data (or input data) may contain errors, and biases as a result of improper data collection practices. The threshold of accuracy required depends on the operational context of the algorithm, as more is not always necessarily better. Current research on the accuracy of algorithms for justice and security has been mixed. Appropriately accurate algorithms can add value to justice and security, however inaccurate algorithms (in conjunction with poor deployment and data management practices) can create inefficiencies, facilitate pernicious feedback loops and can even endanger life and liberty of their direct or indirect targets. What stands out for the value of accuracy, is the necessity of choosing the correct inputs and assuring data quality, predictive validity and eventually real-world outcomes—algorithms must be investigated, scrutinised and tested carefully.

## 2.2 Autonomy

Broadly speaking, a popular conception of autonomy is that of self-rule or self-government (May 1994; Darwall 2006). To live a life freely envisioning one's version of 'the good life', implementing one's decisions and pursuing one's goals without undue constraints and influence has obvious appeal.

[5] COMPAS "…predicts a defendant's risk of committing a misdemeanour or felony within 2 years of assessment from 137 features about an individual and the individual's past criminal record" (Dressel and Farid 2018, 1).

[6] It should be noted that separate by Northpointe (as then known), argued that the predictive validity of COMPAS was acceptable (see Brennan et al. 2009).

[7] This has implications for privacy too. We expect that data be used only as necessity demands, however necessity cannot demand it if the required variables for calculation do not add value in terms of accuracy.

The perspective adopted here is one which is cognisant of the fact that humans are subject to external influences that do not necessarily diminish their autonomy—autonomy is present where an agent has access to their evaluative faculties and their "determinations" are not dictated by circumstance (May 1994, 141). Illustrating this view of autonomy, May (1994, 141) argues:

[a]utonomy does not require detachment from external influences. Rather, it requires that the agent actively assess these influences rather than simply react to them. External influences do not cause action, but rather provide information that the agent, as "helmsman," then steers according to… What we mean when we say a person has autonomy is that she does not simply react to her environment and other influences, but actively shapes her behavior in the context of them.

More succinctly, two keys properties of autonomy are intentional action and conscious reflection (Hildebrandt 2008, 27).

Algorithms pose an interesting problem for autonomy due to their perceived authority, whether justified or not. To some extent they substitute for human cognitive activities and are trusted artefacts which may lead agents to uncritically act on their suggestions, as we have demonstrated above, such trust may not always be merited. As argued by Amoore and Goede (2005, 150) "…questionable data become hardened facts." This authority and uncritical trust borne by algorithms is arguably a threat to autonomy given the right combination of circumstances. This threat may manifest where this influence is reacted to, rather than critically assessed by an agent, bypassing an appropriate evaluative assessment (May 1994). Agents may trust the judgement of an algorithm above their own (Introna and Wood 2004, 14), and in certain contexts algorithmic decisions will automate the responses of their users (Amoore 2011, 38).

Van den Hoven (1998, 97–108) has been critical of the potential influence of IT artefacts in professional situations, positing that agents can be narrowly embedded (or maximally, epistemically enslaved) in epistemic niches supported by software (in our case, algorithms) systems under conditions of inscrutinizability[8] (of the system), pressure, error, and absence of discursive scrutiny. Van den Hoven (1998, 103) argues that in such situations the system output imposes itself on the agent to "….carry their own recommendation as valid, accurate and worthy of belief," and "[c]ertain propositions by the artificial authorities carry themselves as coercive facts." Such a position is an evolution of long held arguments, the likes of which we can see at least as early as 1977 when Joseph Weizenbaum (1977, 236) warned of humankind's increased reliance on the decision-making capacity of computer systems that could not even be understood.

Of this danger, van den Hoven (1998, 104) argues:

> ...one can foresee that by exposing oneself to an epistemic niche, a system's environment or a computer model of a particular part of the world, that part of the world will come to appear as it is rendered by the epistemic artefact in question.

Where the agent is epistemically enslaved, van den Hoven (1998, 105) argues that non-compliance with the system's output can be a form of moral risk taking, where the agent can provide no moral justification. This is a strong thesis, and one which was later largely moderated by Rooksby (2009), arguing that agents are not compelled to believe or act on system output. The listed conditions are nevertheless persuasive, particularly as we imagine a beat officer patrolling a flagged high-risk neighbourhood, or one who meets an SSL subject whilst on patrol (Ferguson 2017b, 79). They may be suspicious, and afraid, and may have no known reason to doubt the validity of their information. Trust and ignorance may obscure reality. They will be free to act but those actions may be constrained by what they deem to be a logical course of action based on what is *to them* factual information (Ferguson 2017b, 85). It might be that the human agent reacts to the algorithm's influence, rather than "actively assess" it (May 1994, 141), and such active assessments are constrained by the conditions listed by van den Hoven.

Ferguson (2017b, 97, 136) confirms the opacity problem faced by agents using information artefacts, arguing that they have no way to check or verify their data and by design must defer to it. Additionally, he seems to indicate that the pressure condition is to some extent present, as some officers will act on information without checking for contextual information from peers and records (this however is notably an active and, strictly depending on the context, potentially a negligent choice) (Ferguson 2017b, 97). The problem is that between the opacity of algorithms that may in truth be difficult to assess and evaluate as legitimate sources of factual information, and other external forces, the kind of evaluative assessment necessary for autonomy may not be possible.

It would seem that whilst algorithmic output does not necessarily make our choices for us, there is a danger that they influence our actions in a way that is at odds with our autonomy and removes us from satisfactory control of our actions and decisions. They are a filter on the lens through which we see the world, and when our vision of the world

[8] Or, in more modern parlance, opacity—the systems, processes and their reasoning cannot always be seen, inspected, or evaluated by their users.

is altered, how we interact with it is constrained by how we see it. A distorted view of the world may lead to distorted choices.

In a firm example of limited human discretion at odds with autonomy, Eubanks (2018) describes an algorithmic system called the Allegheny Family Screening Tool (AFST)[9] that predicts children at risk of abuse. Whilst it is apparent that some discretion is exercised and call centre workers of Allegheny County Office of Children, Youth and Families (CYF) do not automatically defer to the data, they also cannot avert an investigation of a family if a risk score is calculated beyond a certain threshold (20) without a supervisor's intervention, despite the algorithm being "….routinely wrong about individual cases" (Eubanks 2018, 141–142). This example highlights that organisational procedure, that is, the rules governing implementation and deployment of an algorithm, can constrain individual autonomy in potentially undesirable ways. This example shows that whilst one may also be able to actively assess influences of their environment, their determinations may still be subordinated by an algorithm's decision and effectively narrow the waters in which they can sail the proverbial ship. Perhaps in this case, autonomy is not lost so much as it is undermined and devalued.

Where decision-makers (judges for example) hold power over the freedom of those subject (defendants) to algorithmic analysis, we can see that their potentially constrained autonomy may have severe consequences for those subjects. ProPublica reports the case of a criminal who reached a plea-deal for a minor offence involving the theft of a lawnmower and tools, with his prosecutor recommending a year in jail and follow-up supervision thereafter (Angwin et al. 2016). The judge presiding over the case dismissed the prosecutor's recommendation on the basis of the defendant's high COMPAS risk assessment score, and effectively delivered a sentence twice as severe as that recommended by prosecution, stating ' "[w]hen I look at the risk assessment…it is about as bad as it could be"' (Angwin et al. 2016). Here, the influence of the algorithm on the judge may have contributed to the more complete and prolonged reduction of the convict's autonomy. On appeal, after testimony by one of COMPAS' original creators, the judge reversed the harsher sentence, and stated "[h]ad I not had the COMPAS, I believe it would likely be that I would have given one year, six months [a shorter sentence]" (Angwin et al. 2016). It is difficult to determine with confidence here whether the judge's autonomy was compromised in this case, or to what extent. He was in a position to gather a great deal of information pertaining to the case, and presumably the algorithm (it was with thanks to a testimony by one its creators that the decision would be reversed) (Angwin et al. 2016). Nevertheless, in the first instance, the judge deferred to the instrumental rationality of COMPAS, and having later reversed his decision it is evident that he did not actively assess the algorithm's influence on him, and let it subordinate his own practical wisdom.

Practical wisdom is important to consider here, as it is closely related to autonomy (see May 1994, 139–140), it can only thrive in (and is characteristic of) autonomous individuals and is necessary to make fitting moral decisions. Practical wisdom, a concept most notably developed by Aristotle (2004), denotes a certain experiential knowledge of the good and the right, and capacity for reasoned decision. It is learned from experience (not automatically endowed), action and observation and amounts to an ability to recognise the most morally salient features of a situation and act accordingly (see Hursthouse and Pettigrove 2016). It is a product of the correct recognition and application of virtue in the virtue ethics tradition,[10] but care, empathy and openness to understanding and learning are intuitively of great importance, for "[t]he virtuous person not only tends to think and act rightly, but also to feel and want rightly" (Vallor 2018, p. 18). Tapping into our practical wisdom implies appropriate evaluative, reflective capacity and the opportunity for rational choice. Situations of epistemic enslavement are contra to the requirements of practicing practical wisdom, which may be at the risk of being subordinated instead by the instrumental rationality of algorithms, under less than ideal conditions that support the autonomy of human decision-makers. Machines are not capable of such knowledge and decision (see Weizenbaum 1977, 208, 227), and can at best only hope to supplement our evaluations with helpful insights—"[c]omputer systems do not admit of exercises of imagination that may ultimately lead to authentic human judgement" (Weizenbaum 1977, 240). The algorithm will issue decisions without the care, empathy, or imagination that humans are capable of—they may not account for unexpected variables, such as an actively rehabilitating convict standing trial, who may be condemned solely by aspects of his or her upbringing and history. They rely on instrumental rationality and will usually treat problems as technical. As similarly argued by Kitchin (2017, 11) on the topic of city analytics, "…instrumental rationality should not be allowed to simply trump reason and experience, and other sources of information and insight…".

Algorithms can only function effectively where they are teamed with human decision-makers who understand their

[9] This example is not formally within the domain of justice and security, though the link to child safety makes it philosophically relevant to this research.

[10] However, practical wisdom is not exclusive to the virtue ethics tradition (see Audi 2005).

limitations, and are sufficiently free to ultimately rely on their own practical wisdom—computers and algorithms decide, but only a human can choose (Weizenbaum 1977, 259).

More threats broadly linked to autonomy exist, such as lack of consent to data processing and the impact of the generation of derivative data on an individual's identity and self-determination. These will be explored in the following sub-section, privacy.

It might be noted that in low pressure environments, where a plurality of information (and sources) exists and algorithm users are aware of its processes, inputs and limitations, they may in fact enhance autonomy by providing their users with more information to put to use in informed action. Those involved in criminal and security research and strategic planning are perhaps more likely to be less constrained by algorithms, with the time and information available to use (and evaluate) the algorithm to help them understand and explain phenomena, and determine appropriate responses to related problems. Those operating in more real-time or tactical contexts may not have the same access to relevant information, or the time (or authority as we have seen) available to effectively exercise their autonomy.

Borrowing May's (1994) metaphor of a helmsman steering a ship, by providing useful and reliable insights, algorithms may help navigate the ship just as a compass (not COMPAS) might, rather than force it towards the shore like a powerful gust. The conditions of their design, implementation, and deployment must support this, however, instrumental reason should not subordinate practical wisdom.

Thus far, we have examined the concept of autonomy vis-à-vis decision makers and agents of justice and security, but have not yet discussed subjects of algorithmic output. Of course they are of consequence for them too, as we have already seen, but there are some points which warrant elaboration here.

From the perspective of the subjects of algorithms (either as individuals or individuals living within certain areas of interest), algorithms tend towards sewing suspicion and scrutiny that can create adverse contacts with the justice and security system, or even (as in the case of COMPAS) foreclose future opportunities and freedoms. When actions are taken based on algorithms, the will of agents of justice and security may be imposed on those subjects in ways that are at odds with their autonomy, their own capacity for intentional action and conscious reflection. Whether this takes the form of stopping that subject on the street, or delivering a particularly harsh prison sentence, that subject's world will seem smaller and their capacity to make choices and interact with it will be affected to a smaller or greater degree. Mayer-Schonberger and Cukier (2013, 162) warn that predictive algorithms (like COMPAS) could erode the presumption of innocence and deny personal responsibility and accountability by punishing individuals before they have committed a crime. In this situation the individual would have been denied the opportunity of autonomous action, and to later face just punishment for that action.

However, not all interferences with autonomy are necessarily unethical should they represent a regulation of possible harms, as suggested J.S. Mill's harm principle (Brink 2018). Whether an intervention against an individual is a justifiable interference with their autonomy is rather dependent on the nature of that intervention, whether it is proportionate and necessary (for upholding relevant moral values, or the rights of others). Again, relying solely on a COMPAS or similar risk-assessment algorithm and deeming an individual a likely recidivist warranting additional prison time would be an extreme and arguably unreasonable case. On the other hand, inviting an individual or offering him or her a voluntary visit with a social worker[11] would be a minimal and arguably innocuous interference at worst. Losing one's opportunity to make a moral choice to commit a crime is not equivalent to disvalue, however disproportionate means being utilised to foreclose that opportunity would (for example, in the extreme, pre-emptive imprisonment).

It is worth noting that broader societal interventions involving social, infrastructural support combined with law enforcement intervention have shown success in New Orleans, indicating that algorithms that can identify places and people at risk can be perhaps most fruitfully used when not exclusively in the domain of justice and security, but in a more encompassing and collaborative context (Ferguson 2017b, 40–42).

The deployment of algorithms by justice and security agencies needs to strike a fair balance between individual autonomy and the need for safety (see again Brink 2018). The deepening of suspicion of algorithmic targets, as well as increased surveillance entailed by the design and implementation of algorithms and systems that draw data across institutional boundaries may cause a chilling effect and avoidance of institutions that provide vital social goods, and thus constrain autonomy (Brayne 2017, 997–999). We will discuss this also in more specific detail below.

In sum, autonomy may be constrained or undermined by the pressure, opacity, and the perceived authority of algorithms that may preclude a decision-making agent's ability to critically assess it and properly make reasoned decisions (thus also impacting their exercise of practical wisdom), or institutional rules that empower algorithms to more

[11] As regards the SSL, one intervention used according to Ferguson (2017b, 38) is a custom notification visit involving members of the community, a police officer, and a social worker, where a letter of warning is handed over to the SSL subject. Deterrence is the primary motivation of such visits, rather than support, and as such may not be so innocuous.

actively dictate the actions of human agents. Such outputs and actions, as they apply to algorithmic (data) subjects, may also rather more tangibly impact autonomy where they causally contribute to the outright loss of liberty and therefore the exercise of autonomy of those subjects.

## 2.3 Privacy

Privacy is principally (if reductively) defined normatively as "the right to be let alone", and usually encompasses ideas of control of and access to our physical space and personal information (in its many forms) (Warren and Brandeis 1890, 205; Moor 1997; Solove 2005; Tavani 2007; Nissenbaum 2009; Floridi 2013, 228–260; Koops et al. 2017). We are primarily concerned here with our informational privacy, though as data can be generated and processed from many analogue domains, we also accept that this particular variant of privacy overlaps with many other types of privacy (see Moor 1997; Tavani 2007; Koops et al. 2017).[12]

Privacy serves important purposes, allowing us to think and communicate with some qualified freedom, to form relationships, to manage and form our identities without undue interference, to participate in politics without fear (casting a vote), and to protect our safety (from stalkers, for instance) (Nissenbaum 2009, 75–88). Privacy has an important association with autonomy, and unjustified interferences such as egregious surveillance may chill our actions (Solove 2005; Penney 2016). For example, if we are aware of internet surveillance by the state, it may alter the kinds of content that we view online (Penney 2016).

Here, we reference Nissenbaum's (2009) privacy as contextual integrity of information (CI), which proposes that privacy is the right to the appropriate flow of personal information, that is, privacy is respected when our personal information flows in a manner that adheres to the norms of a given context (or for one example, a relationship such as police-officer and crime-victim), where our attributes (types of information) are transmitted by appropriate actors (perhaps police department and prosecutor as an extension of the initial example), under appropriate principles (such as consent, or likely, need) (Nissenbaum 2009, 129–157). Such an account of privacy emphasises its relational nature, which is to say that the flow of information in contexts is determined by norms and properties of those contexts and the relationships between those properties (for example, people and institutions). The theory generally posits boundaries between contexts that are not readily collapsible, and contextual norms that are not readily transferrable. Where deviations from entrenched norms occur in novel practices, a red flag is raised, and warrants some dialogue and reflection (Nissenbaum 2009, 129–157).

Algorithms are often profiling technologies, and both potentially require potentially personal information and generate it. Hildebrandt (2008, 19) provides a useful definition of profiling in the Big Data context that should be noted:

> The process of 'discovering' correlations between data in databases that can be used to identify and represent a human or nonhuman subject (individual or group) and/or the application of profiles (sets of correlated data) to individuate and represent a subject or to identify a subject as a member of a group or category.

Hence, the very construction of a profile indicates the possible generation (or prediction) of personal (see Crawford and Schultz 2014, 98) or demographic information within unexpected contexts, and raises questions of the origin or source of the data (which may not even itself be PII (Crawford and Schultz 2014, 101)) used, and how such profiles are acted upon.

A first concern here is a potential disregard for consent in the migration of data from one database or platform to another, or the movement of personal information from one context (for example, social media) to another (a police database), in order to support or facilitate algorithmic analysis. The advent of Big Data in general has been largely undermining the principle of consent, where the volunteered information of a small number of people can still generate information about those who do not consent to data collection and processing, either through inferences on shared traits or, as we have seen from recent Facebook controversies, from data accessed about persons in one's network for which no consent has been offered (making your friend something of a Trojan horse of data collection) (Barocas and Nissenbaum 2013; Hautala 2018). This is most apparent in the US with the advent of fusion centres that combine inter-agency data about individuals, as well as the collection or purchase of personal information from commercial contexts or scraping of social media data to establish associational networks (Privacy International, nd; Crawford and Schultz 2014, 104; Brayne 2017, 993; Ferguson 2017b, 2, 15; Winston 2018).[13] Beyond standard algorithmic deployments in local or national policing, there is also the much more advanced and potentially encompassing and penetrating capture of

[12] Koops et al. (2017, 566) have produced a comprehensive multidimensional typology of privacy including bodily privacy and intellectual privacy (in the personal zone), spatial and decisional privacy (in the intimate zone), communicational and associational privacy (in the semi-private zone) and proprietary and behavioural privacy (in the public zone).

[13] A further interesting example of context leak occurs in a report by Garvie et al. (2016), who found that 26 US states can deploy facial recognition algorithms in driver's licence databases, and furthermore that the network that such algorithms have access to images of 117 million Americans.

information for surveillance and algorithmic sorting by national security agencies such as the National Security Agency (NSA) (see van der Velden 2015) that represents a more extreme collapse of independent contextual spheres of privacy.

It is not merely criminals nor suspects who may be swept up in data collection practices, as Brayne (2017, 992) observes, police (in the US) are increasingly using data on persons with no prior police contacts. Brayne (2017, 992, 994) offers the interesting example of network analysis as offered by the Palantir platform, which has access to disparate data sources. The Palantir network analysis shows associational webs of entities relating to a person who has had prior police contact, including people and vehicles or phones (Brayne 2017, 992). Some of those persons appearing in this web have not had prior contacts with the police and are included in a database simply by association, and may be colleagues or family (Brayne 2017, 992). Brayne (2017, 992) calls this a network of non-suspect/criminal persons a secondary surveillance network. Brayne (2017, 998) argues that such surveillance will disproportionately impact minorities and persons in poorer neighbourhoods (particularly we might imagine where predictive geo-spatial risk algorithms are also being utilised). Such secondary surveillance networks represent a potentially unjustifiable interference with privacy as we understand it—we would not expect innocuous personal connections to be documented in a police database without consent or clear need.

The lack of consent in data processing in the domain of justice and security is not necessarily always wrong, even if it is an apparent interference with our autonomy. We expect governance institutions to act somewhat coercively in providing law and order, and it is in fact a norm for them to interfere with our privacy rights through data collection and processing when this is lawful, necessary and proportionate—though it is also a norm for them to only use a particular method if it is the least intrusive available (see generally the human rights scholarship of Fox-Decent and Criddle 2009; Fox-Decent 2011; Criddle and Fox-Decent 2012). The context of justice and security does have legitimate and justifiable, exceptional reach. However, indiscriminate and/or large-scale processing of personal data from disparate sources and contexts would be difficult to justify in most scenarios and it is not a norm we would desire to be entrenched.

Another issue stemming from algorithmic analysis of disparate data is its transformative and derivate or generative potential. Algorithms can learn (or predict) new data about you from the data provided to them, such as with Chicago's SSL. This also has striking implications for autonomy and an individual's development of a personal identity, potentially assigning them data on which they will be judged, which will influence their interactions with police, and as we have seen, may not always be an accurate reflection of who they are. This categorisation of persons based on shared characteristics or traits that we see in something like the SSL is interesting in the sense that it results in classifications or categories, or groups, of people (see Floridi 2017; Kammourieh et al. 2017 for more on group privacy) whose interactions with police may differ from everyone else's where they may be observed with more suspicion from or come into more contact with the police. In such cases we see groups being designed by data scientists and government agencies, and encompassing potentially unaware individuals, whose interactions with the state may be adversely affected, thus potentially creating a violation of the nascent concept of group privacy (see Floridi 2017; Mantelero 2017, 145). Such groups can be categorised into binaries of deviants and conformers (de Laat 2019, 5), and there will be the risk of the stigmatization and marginalization of deviant categories (Harcourt 2005, 36–37). These "deviant" categories will naturally be subject of suspicion and scrutiny, and they may disproportionately overlap with minority or marginalized groups, which seems particularly a risk if we recall the pernicious feedback loop (Amoore and De Goede 2005; Guzik 2009, 12). So harm arises as a result of a breach of privacy, and one which is rather discriminative. Again, we have the generation of data entering a context without consent, and with need the requires justification.

The volume of data being processed also raises questions. As indicated by Dressel and Farid (2018), more features may not always equate with more accuracy. If features correspond with personal data and their presence in a data set does not improve the accuracy of an algorithm they cannot fairly be said to be necessary, and their storage, transmission, and processing in new contexts may not deemed proportionate or easily justified.

Untrammelled data collection and processing into algorithmic output represents a threat to privacy and one which may only be prevented or challenged through the acquisition of knowledge regarding the algorithms input data, the terms of their collection, their sources of origin and the practices that underlie them, and the accuracy of the algorithm itself. This point outlines the importance of accountability and transparency, and the necessity of Big Data due process (Crawford and Schultz 2014).

In sum, viewing privacy as the right to an appropriate flow of information, we can see risks arising from the movement of personal data between contexts for the benefit of an algorithm, that may be disproportionate and not easily justified, as well as the creation and categorization of groups (and quite probably overlapping with disadvantaged groups or minorities) that may be more likely to have adverse experiences with agents of justice and security.

## 2.4 Fairness and equality

We expect to be treated with equal regard to our fellow citizens regardless of personal circumstances or characteristics, by other private citizens, organisations and statutory institutions, generally on the basis of the inherent dignity associated with our humanity. Actions taken based on arbitrary (in the sense of unjustified) distinctions with regards to our personal characteristics (such as ethnicity or gender) may constitute discrimination—a difference in treatment between two different persons in relevantly similar situations—and runs against our expectations of equality (Harris et al. 2009, 579; Fox-Decent 2011). Arbitrary is a key term here—while we expect to be treated with equal regard and concern, this does not preclude positive discrimination that promotes the interests of disadvantaged members of society (Harris et al. 2009, 611; Fox-Decent 2011), thus, a distinction can be made between fair treatment and equal treatment.[14]

An enduring normative formulation of equality and fairness comes in Rawls' Principles of Justice, the first of which is particularly relevant here:

> First Principle: each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all (Wenar 2017 citing Rawls 1999)[15]

For this adequate scheme of equal basic liberties to obtain, those who are disadvantaged or marginalised and generally overlooked by a status quo that favours and entrenches the power, privilege and perspective of the majority (or even elite) warrant particular consideration in the development of policy and methods and tools in justice and security (see D'Ignazio and Klein 2018, draft, for extensive discussion along these lines). Indeed, according to Rawls' difference (or second) principle (1999 as cited by Wenar 2017), any social or economic inequalities that do exist "… are to be to the greatest benefit of the least-advantaged members of society." Fairness precludes any policy or action that would perpetuate inequality, or arbitrary unequal treatment. We can be understood to firstly have a negative duty not to cause inequality and discrimination, and a positive one to prevent it.

As stated earlier, algorithms are not necessarily accurate and objective artefacts—they may reflect their creators' biases or bias found in source data (O'Neil 2016, 25). Algorithms imbued with bias can have destructive impacts, manifesting into discrimination against minorities and disadvantaged groups (Barocas and Selbst 2016), thereby violating, for instance, Rawls' first principle of justice.

Barocas and Selbst (2016) exhaustively trace the origins and potential impacts of bias in algorithms. The problem might find its roots at the early stages of the development of an appropriate model, for instance during problem specification where data scientists are required to define target variables (outcomes of interest) and class variables (categories) (Barocas and Selbst 2016, 678). Problems arise during the specification of a target variable and problem in formal terms that are understandable to computers, and "[t]hrough this necessarily subjective process of translation, data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes" (Barocas and Selbst 2016, 678).

This stage is rather important, as it forms the foundation of how an algorithm sees the world and what it has to say about it. Important decisions are made about the nature of the world by a limited number of privileged people, possibly as informed (but not necessarily "dictated") by client institutions (the police etc.) (de Laat 2019, 4), without a sufficient understanding of the circumstances or lived experience of those whose everyday lives their algorithm will ultimately affect (D'Ignazio and Klein 2018, draft). A decision about a classification as *ostensibly* simple as gender can have an ultimately unfair consequence, taking an example offered by D'Ignazio and Klein (2018, draft) in their instructive work, *Data feminism*:

> No one but a gender non-conforming person would know that, before you step into a scanning machine, the TSA agent operating the machine looks you up and down, decides whether you are male or female, and then pushes a button to select the appropriate gender on the scanner's touch-screen interface. That decision loads the algorithmic profile for either male bodies or female ones, against which your measurements are compared. If your body's measurements diverge from the statistical norm of that gender's body—whether the discrepancy is because you're concealing a deadly weapon, or because the TSA agent just made the wrong choice—you trigger a "risk alert," and are subjected to the same full-body pat-down as a potential terrorist.

This example, quoted at length, demonstrates how easy it may be for one to fall into a so-called "deviant" category, not through any fault of their own, but because of decisions

[14] As argued by Evan Fox-Decent (2011, 183), "…equality does not mean equal treatment, it means equal concern and respect. Relevant differences between two persons or their circumstances can justify differential treatment on grounds of fairness".

[15] Of course, also of importance is the second principle, also indicating the justice in positive discrimination via the difference principle:

> Second Principle: social and economic inequalities are to satisfy two conditions:
> They are to be attached to offices and positions open to all under conditions of fair equality of opportunity;
> They are to be to the greatest benefit of the least-advantaged members of society (the difference principle) (Wenar 2017, citing Rawls 1999).

made about the nature of the world made without sufficient concern or knowledge of their impacts. This blindness is exacerbated in other ways. Often data is not collected about minorities or problems commonly and particularly experienced by them, their "…bodies go uncounted…" (D'Ignazio and Klein 2018, draft). Because of this failure to mobilise data about and for the underrepresented, algorithms also cannot be developed that might create actionable insights for these problems (for instance, as indicated by D'Ignazio and Klein (2018, draft), police killings, and maternal health).

Another issue, as stated earlier, is that algorithms learn based on the data which they are trained on. Barocas and Selbst (2016, 680) further argue this, and explain that when instances of prejudice and discrimination in training data are treated as valid examples the algorithm will simply perpetuate these injustices in its outputs, and similarly that where an algorithm's training data represents a biased sample of the population, decisions made based on the algorithm may disadvantage minority populations. Furthermore, removing variables correlated with minority groups (even proxy variables, that is, variables disproportionately attributable to particular groups), may compromise the overall accuracy of an algorithm, rendering solutions difficult, and it may in fact be necessary to reduce the overall accuracy in order to prevent disproportionate impacts on minority groups (Barocas and Selbst 2016, 720–722). As highlighted by Ignazio and Klein (2018, draft) and reported by Angwin et al. (2016),[16] among the questions asked in a COMPAS survey to be given to the subject, are questions related to who raised the person surveyed and whether their parents were separated. These question can be used as a proxy for race. Ignazio and Klein (2018, draft) indicate that in the US a majority of Black children grow up in single parent households. Fairness would dictate excluding these data points, however there could be a decrease in overall accuracy (although that is unknown in this particular case).

Data collection in the context of justice and security, with particular regards to policing, is especially noteworthy. If algorithmic models reflect the data they are built on, if they are based on policing data that is racially skewed (minorities may be overrepresented in police databases), then their output too will be racially skewed and motivate further discriminatory practices (Ferguson 2017b, 73; Richardson et al. 2019), the pernicious feedback loop obtains. In this case the bias becomes sedimented in our IT artefact, potentially perpetuating (if not giving a false glean of legitimacy) to biases that manifest as discrimination that may lead to more contacts and adverse experiences between minorities and other disadvantaged groups, and agents of justice and security (Niculescu Dinca 2016, 140–142; Ferguson 2017b, 78, 92; Richardson et al. 2019). In more concrete terms, Richardson et al. (2019, 14) note that 56% of Black men under 30 in Chicago have risk scores on the SSL—a result potentially a product of racially discriminatory policing and also potentially leading to more severe charges upon arrest by the police.

In the case of COMPAS, ProPublica found a clear distinction between results for Black and White defendants, with Black defendants being found 77% more likely to be flagged as higher risk of committing violent crimes in the future, and 45% more likely to commit any kind of crime (Angwin et al. 2016). Additionally, Black defendants were incorrectly predicted to reoffend at twice the rate as White defendants (44.9% versus 23.5%), and White defendants who did reoffend were predicted to not do so at twice the rate as Black defendants (47.7% versus 28%) (Angwin et al. 2016). These findings were however a departure from the research results of the tool's creators.[17] Whether or not COMPAS' outputs manifest in differential treatment between people of different race and ethnicity in the court room or throughout the different stages of criminal justice is unknown.

In another illustrative case, that of AFST (see above), Eubanks (2018, 156–167) argues that some of its variables are proxies for poverty; the variables focus disproportionately on the poor and working class and minorities, and as such represents a kind of poverty profiling that scrutinises these categories of people based not on their actual behaviour, but group membership "…[b]ecause the model confuses parenting while poor with poor parenting….". When a model's world view is skewed in such a fashion, it turns a spotlight on disadvantaged groups that more advantaged groups may escape, and such bias may be enacted as discrimination if these families are targeted for investigation by CYF in potentially disproportionate numbers.

It is apparent that the risk of poorly designed, biased algorithms can result in increased contacts between statutory and judicial agents and minorities and disadvantaged groups, and consequences may be severe. This is not consistent with a reasonable view of equal basic liberties. Addressing these problems, however, may be a significant challenge requiring an innovative approach, because, as stated, removing variables from data sets that correlate with other sensitive variables may reduce an algorithm's accuracy (Moses and Chan 2018, 811). Furthermore, forming technical solutions to bias and discrimination may be a significant challenge as translating theory into design practice is complex—Binns

[16] https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE

[17] Research conducted by Northpointe (now equivant) reports similar predictive validity results for White and Black defendants (Brennan et al. 2009, 31), suggesting no significant disparity between results for Black and White defendants.

(2018, 9) argues that "…a contextually appropriate approach to fairness which truly captures the essence of the relevant philosophical points may hinge on factors which are not typically present in the data available in situ." Understanding the nature of discrimination requires philosophical reflection on the problem, and appropriately addressing this problem in an algorithm's design is not so much exclusively a data science problem, as a social or philosophical one (Binns 2018, 9), and one which invites the views of those often overlooked who stand to be marginalised by an algorithm that contributes to the enactment of unfair practices (D'Ignazio and Klein 2018, draft).

In sum, if we regard fairness and equality as an adequate scheme of equal basic liberties for all (subject to positive discrimination in the name of fairness and not arbitrary discrimination on the basis of personal characteristics), we can see the threats that arise from algorithms. Minorities can be disproportionately adversely affected by algorithms. Fairness/equality stands as a central value in this analysis, bearing a close conceptual and practical relation with the other values examined. Discriminatory practices and limited perspectives can shape inaccurate models that lead to the perpetuation of further discriminatory practices at odds with our sense of fairness and equality, whether that is racially based targeting of individuals or groups for intervention (autonomy) or scrutiny (privacy).

### 2.5 Property and ownership

In many societies, we value private ownership of our various artefacts (from creative products, to land, buildings and more—our property). Copyright and other proprietary protections are enshrined in law across the world, protecting creators' and software developers' rights to monetarily benefit from, and distribute their products with some degree of control over the use to which their product is put. When we speak of property and ownership here, as indicated, we refer generally to the notion of a bundle of rights (property rights), where A (the owner) has a relation to X (the property) defined by (for example) "…the rights to possess, to use, to manage, to income, to the capital, to security, as well as incidents of transmissibility, absence of terms, the prohibition of harmful use, the liability to execution and its residuary character" (Robaey 2015, 49 citing Honoré 1961). Simpson succinctly explains (2014, 145):

> Ownership of property confers property rights. These consist in the entitlement of the property owner to dispose of her property as she wishes (without breaching others' rights). Property rights also oblige others not to interfere with her property. If someone interferes with her property, two wrongs may be done. First, they wrong her by broaching her property rights. Second, they may wrong her by reducing the value of her property.

Notably, as broadly argued by Robaey (2015), these rights held by property owners have correlated duties (property duties) that demand responsibility. Property owners have the responsibility of designing algorithms that do not undermine the rights of others, and should not unreasonably exercise their assumed property rights to the detriment of others (for instance, obscuring problematic features by asserting blanket secrecy to prevent unauthorised copying of their intellectual property).

In many cases, the algorithms used by agents of justice and security are provided by commercial vendors[18] and their inputs and processes are opaque (Ferguson 2017b, 136) as they are protected by Intellectual Property rights to prevent their duplication or preserve competitive advantage. This serves a reasonably legitimate purpose as it allows the creators to benefit from their labours and/or investments, but the result compounds or causes some issues already discussed and others to be discussed soon, notably that algorithms are not necessarily understood by their users, they cannot be challenged by their targets, the appropriateness and accuracy of their model may not be available for inspection and validation, and the presence of bias consequently may go undetected. As of publication of ProPublica's COMPAS investigation, equivant (formerly Northpointe), its creator, was secretive about weightings, and the same was true of Chicago's SSL (Angwin et al. 2016; Ferguson 2017b, 37).[19]

A benefit of creators/IP holders exercising strict control over their product and restricting access to knowledge of its content and processes is that it might prevent the spread of their technology to bad actors (criminal organisations, and authoritarian regimes with poor human rights' records for example) who may adapt and use it towards malicious ends (Hayes 2018, 265–268). In the wrong hands an algorithm may be a powerful weapon, and ownership may be a powerful tool for stemming the flow of potentially dangerous knowledge.

In sum, ownership and property, viewed as a bundle of rights stemming from the relation of the owner and the property, has value that promotes the owners' ability to profit from their property and control its distribution or use, and places them in a position of responsibility (which can be used for the responsible licencing or dissemination of a technology). However, the level of control possible over their

[18] Some examples include PredPol, Palantir, IBM, Accenture, Information Builders, and Hitachi (Palantir, n.d.; Moses and Chan 2018, 809).

[19] However, the City of Chicago does reveal some of this information on its SSL Dashboard, see https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Dashboard/wgnt-sjgb

property can have the consequence of opacity (hiding the internal workings of an algorithm from end-users or the public more generally).

## 2.6 Accountability and transparency[20]

Here we will define accountability as a form of backward-looking (or passive) responsibility, or "….the (moral) obligation to account for what you did or what happened (and your role in it happening" (van de Poel 2011, 39). This is in contrast to forward-looking (or active) responsibility, which refers to the obligation to see to some state-of-affairs and to ensure discharge of entailed duties, which attach to one's relationship to some object or role (which is our over-arching interest here) (van de Poel and Royakkers 2011, 10–13).

An agent is accountable for some object where they have moral agency (or capacity to act responsibly), have some causal relation to the object, and are suspected of some wrong-doing (van de Poel 2011, 47). The agent may also be assessed on their autonomy and knowledge—whether or not their account and justifications for the object in question are acceptable will then determine if they are to blame (van de Poel 2011, 45–47). Conceptualisations of accountability can vary, however some key points and its instrumental purpose broadly cohere—it is an evaluative process to identify some fault and (if necessary) apportion blame to a causal or responsible agent or agents, and serves a social (or sometimes technical) function of prescription or deterrence from and correction of wrong-doing and fault (see Nissenbaum 1996; Stahl 2006; Floridi 2013, 154–157).

Information is an essential ingredient in accountability. We need information about some event or outcome and about the network of agents and artefacts involved before we can adjudicate who is answerable or what is at fault, and especially who is blameworthy. It may be one agent we hold accountable, or it may be more appropriate to hold several accountable. In our context of algorithms in justice and security, the traceability of fault and causation in particular can be difficult to discern with regards to the so-called "problem of many hands". Many agents may be embedded in a network or chain of events (from design through to implementation and deployment) leading to some problematic outcome, from the agents behind the creation of the algorithm (executives and software engineers or data scientists), policymakers commissioning or authorising the algorithm, to algorithm users (for example, police officers whether on the beat or strategizing behind the scenes in their department) (Nissenbaum 1996, 28–32). Bugs may arise in a software system that could have been unforeseeable and which may strain the application of accountability (Nissenbaum 1996, 32–34). If we recall the problem of epistemic dependence or enslavement, do we hold a police officer who inappropriately fires a gun based on the information he has received from an artefact accountable, or those who designed or maintained the artefact? In this case we can fairly argue everyone involved somehow in the network leading to the harm must answer for it, yet we still need satisfactory information about the event and elements underlying this network to fairly apportion blame.

Opacity, the opposite of transparency, may prevail if certain agents are not forthcoming, or the algorithms they rely on are not accessible or understandable. Transparency we define synthetically here as a state-of-affairs conducive to knowledge about some X (for instance, our algorithm) characterised by availability, accessibility, and understandability/explainability of relevant information (see Heald 2006; Menéndez-Viso 2009; Turilli and Floridi 2009; Tu 2014). Transparency will often be teleological and relational in nature, which is to say normally some X will be rendered transparent to some person(s) Y for purpose Z (whether that is auditing, or informed decision-making, etc.).

Algorithms and the institutions that develop or use them may be black-boxes to the outside world, unseen and not understood. Citing Burrell (2016), Lepri et al. (2018, 619) identify three types of opacity that emerge from algorithms, some of which we already mentioned above. Each type of opacity contributes to a general inscrutability, including of the evidence used (source data) for how an algorithm reaches its determinations (Mittelstadt et al. 2016, 4). The problem with opacity is fundamentally that we cannot see the impact of the algorithm's design, implementation, or deployment on our values—we cannot see data used to determine if our privacy has been violated, or understand its reasoning in order to challenge its decisions.

The first type of opacity is intentional opacity (Lepri et al. 2018, 619), which references secrecy to preserve the intellectual property of the algorithm. In other cases the secrecy of the algorithm is maintained to protect the algorithm's efficacy and its "competitive advantage" where, under the rationale that if the inputs and model are known, it might be vulnerable to manipulation, duplication, or tactical advantage could be lost (Mittelstadt et al. 2016, 6; Ferguson 2017b, 136). Ferguson (2017a, 1187) for one argues that if criminals were aware of social network linkages or predictive targets, they might be sufficiently informed to adapt to or counter intervention strategies. This argument may be a legitimate one, depending on the precise circumstances, but nonetheless may exacerbate application of accountability. Intentional opacity represents a problem of accessibility (excluding access of others) (Mittelstadt et al. 2016, 6). We might add that unintentional opacity is also quite

[20] These are obviously two different values. As they are of functionality very supportive values in this context, they will be discussed together.

possible whereby those responsible for algorithms do not disclose potentially important information about an algorithm's design, implementation, and deployment, simply because they are unaware of the demand for it or necessity of disclosure.

The second type of opacity is illiterate opacity, referencing the lack of requisite technical skills held by most members of society to understand what underpins algorithms. This opacity represents a problem of understandability. This is a low level comprehensibility problem (Mittelstadt et al. 2016)—the algorithm is largely incomprehensible to the general public.

The third type of opacity is intrinsic opacity, and it arises from the lack of interpretability of complex algorithms such as those built around deep learning (Lepri et al. 2018, 619–620). This is more a high level comprehensibility issue, which is to say that an algorithm not only is not generally comprehensible, but may be beyond the capacity of experts to track and comprehend. With regards to machine learning and dynamically changing inputs and outputs, mathematical rather than semantic ontology, as well as the scale and speed of data processing, even expert human oversight may be impeded (Burrell 2016, 2; Mittelstadt et al. 2016, 6).

This opacity surrounding algorithms may make it difficult to ascertain their casual role in some event or state-of-affairs, or even if they were involved in it at all if the veil of secrecy is thick enough that even its existence is unknown (consider the NSA's activities in advance of the Snowden disclosures). Algorithms cannot be effectively challenged or indeed corrected without sufficient knowledge.

D'Ignazio and Klein (2018, draft) strongly emphasise the importance of transparency, not just of the algorithms and data, but of those who create them too. As already stated, those using data may be, in their words, "strangers" to the datasets they use and insufficiently knowledgeable of their context (D'Ignazio and Klein 2018, draft). Those who work on (and perhaps with) algorithms should, by their argument, self-disclose their positions and acknowledge their own privileges and not only that, but seek and invite the standpoints of others, those whose voices are probably not adequately represented in the design process and who are most likely to suffer the adverse consequences of an algorithm's deployment (D'Ignazio and Klein 2018, draft). If transparency is a state-of-affairs conducive to knowledge about some X, then it should also serve the purpose of identifying what knowledge or information is missing, and the situated knowledge of those whose voices often go unheard is a valuable, if not vital, body of knowledge that should be present in institutions designing, implementing, and deploying algorithms. As D'ignazio and Klein (2018, draft) argue:

> The goal of feminist objectivity, then, becomes to connect knowledge back to the bodies of its producers and institutions, with their particular histories, values, limitations, and oversights. In short, to consider context in relation to data.

Thus transparency is important not just for seeing inside and understanding black boxes and their origins and contexts, and implications, but it is also necessary for their designers to see and understand the contexts attached to their data and work.[21] Those working with Big Data must "…pair off quantitative methodologies with qualitative questions" (van Dijck 2014, 206). Such a task may be a collaborative one between designers, agents of justice and security, and algorithmic subjects, particularly those often unheard but potentially disproportionately affected by them.

Transparency (especially when paired with accountability) is a powerful value which conceptually and practically supports many others either directly or indirectly. Knowledge of an algorithm may release agents of justice and security from epistemic dependence and help re-engage practical reason should they appreciate their limitations and underlying context (autonomy). It supports fairness by identifying missing knowledge (like situated knowledge) and situating the perspectives of its creators and users, and can help the design of fairer algorithms. It can help algorithmic subjects understand and challenge decisions made against them, or correct inaccurate data (autonomy and privacy). Generally speaking, transparency and accountability combined allow us to evaluate threats to our values and apply pressure, or appeal to legal machinery, in order to address and prevent or remedy those harms.

Transparency is nonetheless difficult to achieve, and it too in practice may cause disvalue (we would not want de-anonymised or otherwise personally identifiable data transmitted), facilitate algorithmic manipulation or evasion by suspects, or duplication of intellectual property. Transparency, like many values, is not zero sum. The current of much academic thought is that transparency of algorithms can be limited or bounded in certain situations to a kind of "qualified" transparency, where access is provided to internal or external auditors or public bodies who can test them and investigate complaints (Mayer-Schonberger and Cukier 2013, 180–182; Crawford and Schultz 2014, 124; Pasquale 2016, 160–165; de Laat 2018, 534). Organisations would still hold a responsibility to be more widely open about their own properties, standpoints, and practices.

As a final point in this discussion, algorithms provide a particularly notable opportunity for transparency and accountability. According to Ferguson (2017b, 143–166) the power of Big Data could be used to hold police accountable

[21] For further discussion of a similar nature on this topic, see (Ananny and Crawford 2018).

through the collection of data about the police and their behaviours, and used to build algorithms that can predict undesirable behaviours, and reduce the risk of negative police encounters and improve policing generally. Early intervention systems not using Big Data designed to identify and support potentially problematic law enforcement officers have already been adopted to some extent, and promising research utilising machine learning techniques is now being applied to the problem (Carton et al. 2016). In such cases, this data can enhance accountability by creating data about which a police department (for instance) can answer should a police officer be involved in some predicted and avoidable event. From the perspective of forward-looking responsibility, it places a duty on the department to ensure officers identified as being at risk are targeted with appropriate support and training.

In sum, if we understand accountability as the obligation to account for your role in some wrong-doing, in order to apportion blame and to deter further wrong-doing or improve something, we can see that the complex network of artefacts and persons involved in the design, implementation, and deployment of algorithms can make application of accountability (who is answerable and to blame?) challenging on the one hand. On the other hand, the potential of algorithms to be applied to improved performance and behaviour of police (for example) indicates that algorithms can support accountability by creating standards and information by which to hold police to account, and achieve the purpose of accountability by deterring and addressing wrong-doing. We can see that transparency, understood as a state-of-affairs conducive to knowledge about some X characterised by availability, accessibility, and understandability/explainability of relevant information, is inherently challenged by intentional and unintentional opacity that further strains accountability (needed information is simply not available). However, where it is satisfactorily observed, it is possible for algorithmic subjects (or society more generally) to challenge aspects of the algorithm and hold the correct agents to account.

## 3 Disvalue and value relations

Our discussion of seven values that are at play in algorithms in the domain of justice and security demonstrates that these values present challenges at design, implementation, and deployment stages. These values interrelate quite closely and are sensitive to each other (failure to uphold one can often implicate another). Here we will examine the potential of disvalue and potential value conflicts.

Where an algorithm's design and implementation fail to respect a particular value the result may intuitively be the mirror image of what we desire. If an algorithm results in some undesirable state-of-affairs a value is not upheld, but arguably its opposite emerges as a disvalue.

We have seen that algorithms that are not sufficiently (to put it plainly) calibrated with reality will not be faithful to reality in their output—they will be inaccurate. Inaccuracy then is a disvalue. The problem of disvalue here is not discrete and can have recursive consequences. If this inaccuracy is characterised by bias it could translate into discrimination against particular social groups. Here, the value of equality/fairness is replaced by the disvalue of inequality/unfairness as a direct result of the disvalue of inaccuracy.

A failure to uphold one value then may cause a cascade effect, diminishing the possibility for the realisation of other values, and promoting instead the perpetuation of disvalue.

The problem of disvalue and its recursive potential highlights the importance of the conceptual support of related values. Values may also come into tension or conflict in implementation.

As in any domain, values may support each other as well as inhibit each other, or even conflict. When it comes to conflicting and supporting values, one should distinguish between two types of conflict (or support) between values. One is the situation where conflict depends on the contingencies of certain technical solutions or possibilities. One possibility we highlighted is that an increase in fairness could result in a decrease of accuracy (contrarily to the previous example and demonstrating the complexity of the terrain). This particular example depended on the technical solution chosen, which means that attempts can be made to design out the conflict, or compromise on less accuracy if necessary (Van den Hoven et al. 2012; Barocas and Selbst, 2016).

Values may also be conflicting or supportive at a more conceptual level. For example, privacy is often seen as precondition for, or supportive of, autonomy (Kleinig et al. 2011). The point is not that algorithms that protect privacy always also support autonomy, occasionally even the opposite may be true, depending on the specific technical implementation. The point is that the value of privacy is usually seen as a precondition for, and therefore supportive of, the value of autonomy. The argument here is that without a certain degree of privacy, we also lack autonomy, as constant surveillance of what we do and think undermines our capacity to think and decide for ourselves.

If we look at the seven values we distinguished, we find certain as such conceptually supportive relations between the main values. The main ones are summarized in Fig. 1, and elaborated upon in more detail in Appendix A in Supplementary material. At the conceptual level we find supportive but no conflicting relations between the values. Indeed, it seems hard to think of conceptual (or inherent) conflicts between the seven values we discussed. This may be seen as an advantage of the proposed value framework, as

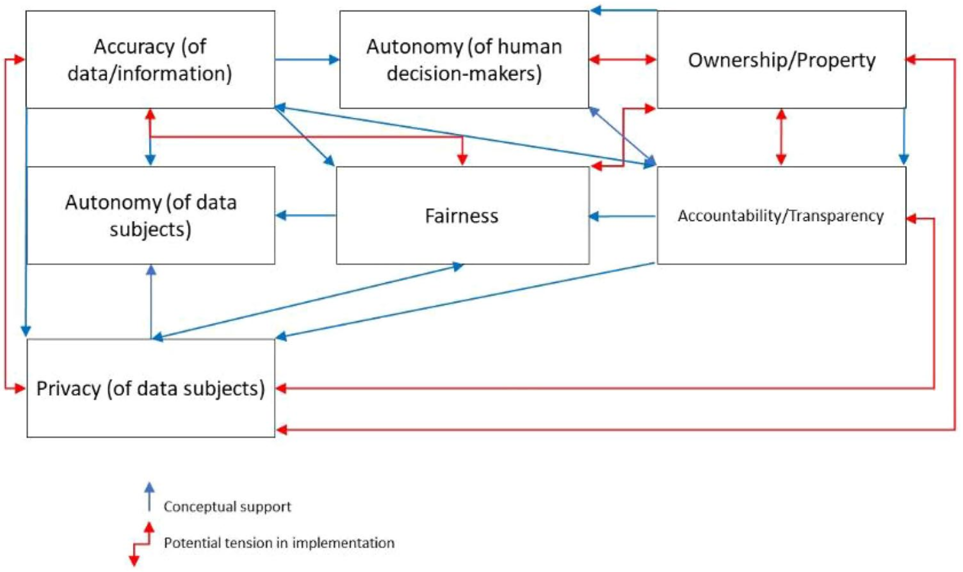
Accuracy (of data/information)
Autonomy (of human decision-makers)
Ownership/Property
Autonomy (of data subjects)
Fairness
Accountability/Transparency
Privacy (of data subjects)
Conceptual support
Potential tension in implementation

**Fig. 1** Conceptual support of values and potential tensions in implementation

it testifies to the conceptual coherence of the framework proposed. This does, of course, not rule out that in the technical and institutional realisation of these values, in a particular context and for a particular algorithm, contingent value conflicts arise. Indeed, current experiences with algorithms suggest not only that sometimes values have been ignored (leading to disvalue) but also some value conflicts that may be hard to overcome. The main potential value conflicts are also highlighted in Fig. 1.

Figure 1 is helpful in better understanding the relations between different values in the value sensitive design of algorithms in the domain of justice and security and in devising design strategies to avoid disvalue and to deal with potential value tensions or conflicts Below, we point out two main ways in which the figure can be used.

### 3.1 Using the figure to better understand the place of single values in the overall framework

One way in which the figure can be used is to better understand the place (and role) of (mostly) single values in the overall framework. We provide three illustrative examples focusing on the conceptual relations between the values:

1. The value of accuracy: as Fig. 1 shows this value provides conceptual support to almost all other values apart from that of ownership/property. Conversely, it is conceptually only supported (directly) by accountability/transparency. In our framework, it is thus a supporting value, i.e. a value that (conceptually) supports other values
2. The value of autonomy (of data subjects): in our framework, this value does not lend conceptual support to any other value. Conversely, it is directly conceptually supported by accuracy, fairness and privacy and indirectly by all values. We may call this a final value.
3. The value of accountability/transparency: as we can see in Fig. 1, this value is conceptually supported by accuracy, autonomy (of human decision makers) and ownership/property. It lends conceptual support (either directly or indirectly) to almost all other values apart from ownership/property. We may call this an intermediary value, in the sense that it requires other values, but also supports several other relevant values in the domain of justice and security.

Our framework thus suggests that we can distinguish between supporting, intermediary and final values. Figure 1 suggests that autonomy (both of data subjects and human decision makers) is a final value, at least in the context of justice and security that we are here considering, while accuracy and ownership/property are supporting values in this context.[22] The other three values (accountability/transparency, fairness and privacy) are intermediary values.

### 3.2 Using the framework to draft design strategies that avoid disvalue and address value conflicts

We may also use the framework to think about design strategies for the value sensitive design of algorithms in the domain of justice and security. We would like to suggest

[22] It should be noted that this claim is dependent on the specific context we are considering (i.e. the domain of justice and security). Moreover, by calling a value supportive, we do not wish to imply that the value has no value in istelf, as values may be both supportive and be valuable in themselves.

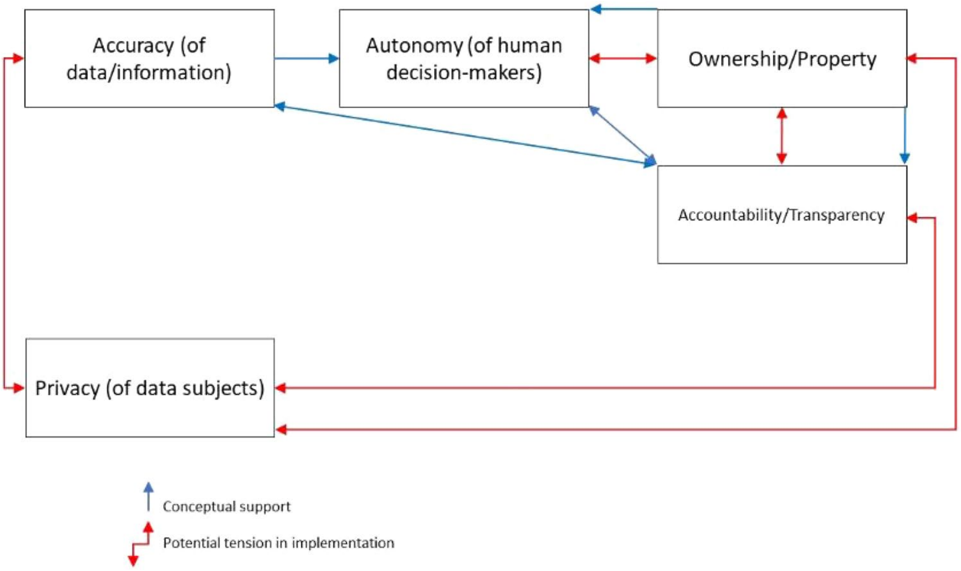
Accuracy (of data/information)
Autonomy (of human decision-makers)
Ownership/Property
Accountability/Transparency
Privacy (of data subjects)
Conceptual support
Potential tension in implementation

**Fig. 2** Accountability/transparency as a target value

that a useful first step is to single out one of two values that one particularly wants to achieve in a particular case. This helps to achieve focus and provides a perspective on which disvalue should in particular be avoided and what value conflicts are most important to consider.

In singling out this value (or these values), three types of considerations are important. First, given that we are looking for values that we particularly want to achieve, the most plausible candidates in our framework are final and intermediary values, not supporting values (which are mainly means to achieve other values). Secondly, our choice will depend on the particular application and context. For example, as highlighted by McCue (2015), what we want from accuracy is highly contingent on context (operational utility) and its implications for those the algorithm will affect (for something such as COMPAS, we expect it to be very high). Third, there may be existing algorithms or an existing non-algorithmic practice that creates disvalue, which we want to avoid in the future.

Once we have selected one or two target values, we may start to look for design strategies with the help of Fig. 1. For example, suppose we have selected accountability/transparency as a target value. Figure 1 then shows us that there are three other values that may lend (conceptual) support to this value, that is, accuracy, autonomy (of decision makers) and ownership/property. See Fig. 2 for a clearer focus on accountability/transparency and its relations.

Each of these three values may provide inspiration to devise more specific design strategies, for example, we may aim to increase transparency/accountability (accountability in particular) by providing more autonomy to the relevant decision makers. This encourages reflection on what strategies, technical or otherwise, can be incorporated to promote autonomy.

In addition, Fig. 1 can be used to identify the main value conflicts we should consider. In the case of transparency/accountability, for example, the figure suggests two main value conflicts to consider, namely the conflicts with ownership/property and that with privacy. The fact that we selected (in this hypothetical example) transparency/accountability as a target value does not imply that we can simply sacrifice these conflicting values (in this case privacy and ownership/property) to achieve more transparency/accountability. The point of our framework is that all seven values are important and should be respected, at least to some minimal degree. Nevertheless, setting a target value may provide a

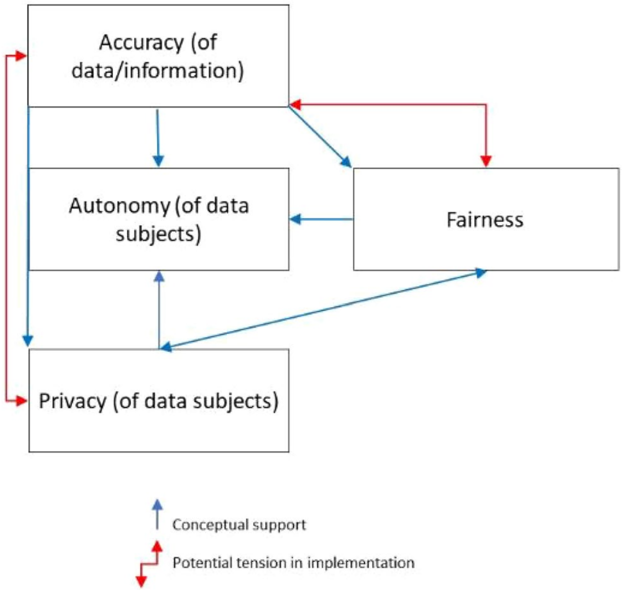
Accuracy (of data/information)
Autonomy (of data subjects)
Fairness
Privacy (of data subjects)
Conceptual support
Potential tension in implementation

**Fig. 3** Focus on privacy

perspective that helps to find ways that respects all values sufficiently, while increasing, if possible, the target value.

Furthermore, in proposing the use of this figure and target values, we must stress that we do not propose ignoring any values. A staged process might be suggested. For instance, after taking accountability/transparency as a target value, designers then might use the figure to select another target value and repeat the process.

What makes the figure particularly helpful when it comes to conflicting values is that it provides a broader context for considering value conflicts. For example, if one chooses to achieve transparency/accountability at the costs of privacy, the figure shows that it is likely that by sacrificing privacy, we may also diminish the autonomy of data subjects and fairness (see Fig. 3). It is this broader context that we should consider in deciding what trade-offs between values are still are acceptable and which ones not. This broader context may also be useful in finding possible win/win design strategies, that is, strategies that serve several values simultaneously.

Finally, there may be nuances in views on values and how they manifest or conflict. Our framework may be subject to some disagreement or modification, but should nonetheless serve as a useful starting point that encourages further discussion and reflection on values in algorithms for justice and security. It is not argued to be absolute, and there could be variations.

## 4 Conclusion

In this paper, we have conceptualised numerous values and from a literature review have examined how these values might fail to manifest in the design, implementation, and deployment of algorithms in justice and security. We find that there is much potential for algorithms to undermine our values, both for end-users and data-subjects. We construct algorithms generally to enhance human autonomy with the end goal of promoting human flourishing, however where there are errors in the process of their design, implementation and deployment, this potential is increasingly constrained. From our brief examination of disvalue, we have highlighted that not only is this potential constrained, but that a failure to adequately incorporate values into the design to deployment process may be actively deleterious to our values and actively inhibit flourishing. We have shown the potential for adverse outcomes from inaccuracy, shown that an uncritical acceptance of algorithms fostered by opacity can undermine autonomy, warned of the potential for gratuitous data collection and processing to undermine privacy, demonstrated the capacity of algorithms to exacerbate discriminatory actions against minorities and other social groups, and shown how ownership and complexity can contribute to opacity and challenge accountability.

All these adverse implications represent problems to be addressed during design, implementation and deployment, and are not insurmountable challenges. This demonstrates the importance of VSD, and incorporating values into the design process. Here, we have provided an initial framework of a set list of values (they are not exhaustive) and conceptualisations that can provide a start to this work.

In order to provide support for the value sensitive design of algorithms, we have proposed a conceptual framework of the conceptual support of values and their practical tension. It is intended that such a framework can help designers reflect on their design decisions, and devise design requirements that can help them uphold values, or at least meet some equilibrium of minimal disvalue.

The challenge of designing algorithms that maximise their contribution to flourishing in the context of justice and security without causing harm is not to be taken lightly, but the rewards are potentially great. Lives and property can be protected if the design, implementation and deployment of algorithms can be executed effectively, and ethically. These challenges are not always going to be possible to solve with mathematical solutions, as some problems require philosophical deliberation as Binns (2018) suggests.

These are problems to be addressed further on, and as required by individual projects. Here, we have presented conceptualisations that can provide a starting point, and indicated the conceptual relations that demonstrate how interwoven values can be.

**Acknowledgements** This research was funded by the Start Impulse Program of the Dutch National Science Agenda (NWA), under The Netherlands Organisation for Scientific Research, VW Data P4 (#400.17.605). The authors are grateful to Remco Boersma, Tjerk Timan, and peer reviewers, for suggestions and feedback.

## References

Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 20(3):973–989. https://doi.org/10.1177/1461444816676645

Amoore L (2011) Data derivatives: on the emergence of a security risk calculus for our times. Theory Cult Soc 28(6):24–43. https://doi.org/10.1177/0263276411417430

Amoore L, De Goede M (2005) Governance, risk and dataveillance in the war on terror. Crime Law Soc Change 43(2):149–173. https://doi.org/10.1007/s10611-005-1717-8

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine Bias, ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 19 Oct 2018

Aristotle (2004) The Nicomachean ethics. New edition. English edition: Tredennick H. (Trans: Thomson JAK). Penguin Classics, NY

Audi R (2005) The good in the right: a theory of intuition and intrinsic value. Princeton University Press, Princeton

Barocas S, Nissenbaum H (2013) Big data's end run around anonymity and consent. In: Privacy, Big Data, and the Public Good: Frameworks for Engagement, pp 44–75. https://doi.org/10.1017/CBO9781107590205.004

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671–732

Binns R (2018) Fairness in machine learning: lessons from political philosophy. J Mach Learn Res 81:1–11

Boateng FD (2018) Crime reporting behavior: do attitudes toward the police matter? J Interpers Violence 33(18):2891–2916. https://doi.org/10.1177/0886260516632356

Brayne S (2017) Big data surveillance: the case of policing. Am Sociol Rev 82(5):977–1008. https://doi.org/10.1177/0003122417725865

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the compas risk and needs assessment system. Crim Justice Behav 36(1):21–40. https://doi.org/10.1177/0093854808326545

Brink D (2018) Mill's moral and political philosophy. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/mill-moral-political/. Accessed 8 Oct 2019

Burrell J (2016) How the machine "thinks": understanding opacity in machine learning algorithms. Big Data Soc 3(1):2053951715622512. https://doi.org/10.1177/2053951715622512

Carton S, Helsby J, Joseph K, Mahmud A, Park Y, Walsh J, Cody C, Patterson C, Haynes L, Ghani R (2016) Identifying police officers at risk of adverse events. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NY, USA: ACM (KDD '16), pp 67–76. https://doi.org/10.1145/2939672.2939698

Crawford K, Schultz J (2014) Big data and due process: toward a framework to redress predictive privacy harms. Boston Coll Law Rev 55:93–128

Criddle EJ, Fox-Decent E (2012) Human rights, emergencies, and the rule of law. Hum Rights Q 34(1):39–87. https://doi.org/10.1353/hrq.2012.0001

D'Ignazio C, Klein L (2018, draft) Data feminism. MIT Press. https://bookbook.pubpub.org/data-feminism. Accessed 17 Sept 2019

Danaher J (2016) The threat of algocracy: reality, resistance and accommodation. Philos Technol 29(3):245–268. https://doi.org/10.1007/s13347-015-0211-1

Darwall S (2006) The value of autonomy and autonomy of the will. Ethics 116(2):263–284

De Laat PB (2018) Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos Technol 31(4):525–541. https://doi.org/10.1007/s13347-017-0293-z

De Laat PB (2019) The disciplinary power of predictive algorithms: a Foucauldian perspective. Ethics Inf Technol. https://doi.org/10.1007/s10676-019-09509-y

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580. https://doi.org/10.1126/sciadv.aao5580

Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York

Ferguson AG (2017a) Is "Big Data" racist? Why policing by data isn't necessarily objective, Ars Technica. https://arstechnica.com/tech-policy/2017/12/is-big-data-racist-why-policing-by-data-isnt-necessarily-objective/. Accessed 26 Mar 2019

Ferguson AG (2017b) Policing predictive policing. Washington Univ Law Rev 94(5):1109–1189

Floridi L (2005) The ontological interpretation of informational privacy. Ethics Inf Technol 7(4):185–200. https://doi.org/10.1007/s10676-006-0001-7

Floridi L (2013) The ethics of information. Oxford University Press, Oxford

Floridi L (2017) Group privacy: a defence and an interpretation. In: Taylor L, Floridi L, Sloot B (eds) Group privacy: new challenges of data technologies. Springer International Publishing, Cham, pp 83–100 https://doi.org/10.1007/978-3-319-46608-8_5 **(Philosophical Studies Series)**

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetege C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI Society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Fox C, Levitin A, Redman T (1994) The notion of data and its quality dimensions. Inf Process Manag 30(1):9–19. https://doi.org/10.1016/0306-4573(94)90020-5

Fox-Decent E (2011) Sovereignty's promise: the state as fiduciary. Oxford University Press, New York **(Oxford Constitutional Theory)**

Fox-Decent E, Criddle EJ (2009) The fiduciary constitution of human rights. Leg Theory 15(4):301–336. https://doi.org/10.1017/S1352325210000017

Friedman B, Khan PH, Borning A (2013) Value sensitive design and information systems. In: Early engagement and new technologies: Opening up the laboratory, pp 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

Friedman B, Khan PH, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta DF (eds) Human-computer interaction and management information systems: foundations, 1st edn. Routledge, New York, pp 348–372

Garvie C, Alvaro B, Frankle J (2016) The perpetual line-up, perpetual line up. https://www.perpetuallineup.org/. Accessed 27 Mar 2019

Guzik K (2009) Discrimination by design: predictive data mining as security practice in the United States' "war on terrorism". Surveill Soc 7(1):3–20. https://doi.org/10.24908/ss.v7i1.3304

Harcourt B (2005) Against prediction: sentencing, policing, and punishing in an actuarial age. Public Law & Legal Theory. https://chicagounbound.uchicago.edu/public_law_and_legal_theory/22

Harris D, O'Boyle M, Bates EP, Buckley C (2009) Harris, O'Boyle & Warbrick: law of the european convention on human rights, 2nd edn. OUP Oxford, Oxford

Hautala L (2018) Facebook kept sharing users' friend data in special deals, report says, CNET. https://www.cnet.com/news/facebook-kept-sharing-users-friend-data-in-special-deals-report-says/. Accessed 1 Mar 2019

Hayes PD (2018) An analysis of emerging ethical and human rights issues in the harvesting of data from social media during emergency response to natural hazards. Thesis. Trinity College Dublin. School of Religions, Theology & Ecumenics. Irish School of Ecumenics. https://www.tara.tcd.ie/handle/2262/82930. Accessed 25 Oct 2018

Heald D (2006) Varieties of transparency. In: Hood C, Heald D (eds) Transparency: the key to better governance? Oxford University Press for The British Academy, Oxford, pp 25–43.https://global.oup.com/academic/product/transparency-the-key-to-better-governance-9780197263839?q=9780197263839&lang=en&cc=gb. Accessed 19 Oct 2018

High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top. Accessed 4 Oct 2019

Hildebrandt M (2008) Defining profiling: a new type of knowledge? In: Hildebrandt M, Gutwirth S (eds) Profiling the European citizen: cross-disciplinary perspectives. Springer, Dordrecht, pp 17–45 https://doi.org/10.1007/978-1-4020-6914-7_2

Honoré T (1961) Ownership. In: Guest AG (ed) Oxford essays in jurisprudence: a collaborative work, 1st edn. Oxford University Press, Oxford

Hursthouse R, Pettigrove G (2016) Virtue ethics. In: Zalta EN (ed) The stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/. Accessed 30 Dec 2019

Introna L, Wood D (2004) Picturing algorithmic surveillance: the politics of facial recognition systems. Surveill Soc. https://doi.org/10.24908/ss.v2i2/3.3373

Kääriäinen J, Sirén R (2011) Trust in the police, generalized trust and reporting crime. Eur J Criminol 8(1):65–81. https://doi.org/10.1177/1477370810376562

Kammourieh L, Baar T, Berens J, Letouzé E, Manske J, Palmer J, Sangokoya D, Vinck P (2017) Group privacy in the age of big data. In: Taylor L, Floridi L, Sloot B (eds) Group privacy: new challenges of data technologies. Springer International Publishing, Cham, pp 37–66 https://doi.org/10.1007/978-3-319-46608-8_3 **(Philosophical Studies Series)**

Kelleher JD, Tierney B (2018) Data science. The MIT Press, Cambridge

Kitchin R (2016) The ethics of smart cities and urban science. Philos Trans R Soc A Math Phys Eng Sci 374(2083):1–15. https://doi.org/10.1098/rsta.2016.0115

Kitchin R (2017) Thinking critically about and researching algorithms. Inf Commun Soc 20(1):14–29. https://doi.org/10.1080/1369118X.2016.1154087

Kleinig J, Mameli P, Miller S, Salane D, Schwartz A (2011) Security and privacy: global standards for ethical identity management in contemporary liberal democratic states. ANU E Press, Acton. http://doi.org/10.22459/SP.12.2011. Accessed 28 Feb 2019

Koops B-J, Newell BC, Timan T, Škorvánek I, Chokrevski T, Galič M (2017) A typology of privacy. Univ Pa J Int Law 38(2):483–575

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 31(4):611–627. https://doi.org/10.1007/s13347-017-0279-x

Lewin J, Wernick M (2015) Chicago police department and predictive policing. International Association of Chief's of Police, Chicago

Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mantelero A (2017) From group privacy to collective privacy: towards a new dimension of privacy and data protection in the big data era. In: Taylor L, Floridi L, Sloot B (eds) Group privacy: new challenges of data technologies. Springer International Publishing, Cham, pp 139–158 https://doi.org/10.1007/978-3-319-46608-8_8 **(Philosophical Studies Series)**

May T (1994) The concept of autonomy. Am Philos Q 31(2):133–144

Mayer-Schonberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work and think. John Murray, London

McCue C (2015) Data mining and predictive analysis: intelligence gathering and crime analysis, 2nd edn. Butterworth-Heinemann, Waltham

Menéndez-Viso A (2009) Black and white transparency: contradictions of a moral metaphor. Ethics Inf Technol 11(2):155–162

Miller S (2009) The moral foundations of social institutions: a philosophical study, 1st edn. Cambridge University Press, Cambridge, New York

Miller D (2017) Justice. https://plato.stanford.edu/archives/fall2017/entries/justice/. Accessed 28 Feb 2019

Mittelstadt BD, Allo P, Taddeo M, Mariarosario T, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):2053951716679679. https://doi.org/10.1177/2053951716679679

Moor JH (1997) Towards a theory of privacy in the information age. SIGCAS Comput Soc 27(3):27–32. https://doi.org/10.1145/270858.270866

Moses LB, Chan J (2018) Algorithmic prediction in policing: assumptions, evaluation, and accountability. Polic Soc 28(7):806–822. https://doi.org/10.1080/10439463.2016.1253695

Niculescu Dinca V (2016) Policing matter(s). Datawyse / Universitaire Pers Maastricht. https://cris.maastrichtuniversity.nl/portal/en/publications/policing-matters(b911f31c-f8e8-44e9-999c-5c8edcafd7b7).html. Accessed 19 Oct 2018

Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2(1):25–42. https://doi.org/10.1007/BF02639315

Nissenbaum H (2009) Privacy in context: technology, policy, and the integrity of social life, 1st edn. Stanford Law Books, Stanford

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy, 1st edn. Crown, New York

Palantir (n.d.) Law enforcement, Palantir. https://palantir.com/solutions/law-enforcement/index.html. Accessed 21 Jan 2019

Pasquale F (2016) The Black Box Society: the secret algorithms that control money and information, Reprint edn. Harvard University Press, Cambridge

Peeters R, Schuilenburg M (2018) Machine justice: governing security through the bureaucracy of algorithms. Inf Polity 23(3):267–280. https://doi.org/10.3233/IP-180074

Penney J (2016) Chilling effects: online surveillance and wikipedia use. Berkeley Technol Law J 31(1):117. https://doi.org/10.15779/Z38SS13

Perry WL, McInnis B, Price CC, Smith S, Hollywood JS (2013) Predictive policing; the role of crime forecasting in law enforcement operations. RR-233-NIJ. RAND. https://www.rand.org/pubs/research_reports/RR233.html. Accessed 20 Dec 2018

Police (UK) (nd) Automatic Number Plate Recognition—Police.uk. https://www.police.uk/information-and-advice/automatic-number-plate-recognition/. Accessed 27 Mar 2019

Privacy International (nd) Social media intelligence, privacy international. https://privacyinternational.org/explainer/55/social-media-intelligence. Accessed 26 Mar 2019

Rawls J (1999) A theory of justice, Revised edn. Harvard University Press, Cambridge

Richardson R, Schultz J, Crawford K (2019) Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. N Y Univ Law Rev 94(2):192–233

Robaey Z (2015) Looking for moral responsibility in ownership: a way to deal with hazards of GMOs. J Agric Environ Ethics 28(1):43–56. https://doi.org/10.1007/s10806-014-9517-8

Rooksby E (2009) How to be a responsible slave: managing the use of expert information systems. Ethics Inf Technol 11(1):81–90. https://doi.org/10.1007/s10676-009-9183-0

Saunders J, Hunt P, Hollywood JS (2016) Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. J Exp Criminol 12(3):347–371. https://doi.org/10.1007/s11292-016-9272-0

Schwartz SH, Bilsky W (1987) Toward a universal psychological structure of human values. J Pers Soc Psychol 53(3):550–562. https://doi.org/10.1037/0022-3514.53.3.550

Simpson TW (2014) The wrong in cyberattacks. In: Floridi L, Taddeo M (eds) The ethics of information warfare. Springer International Publishing, Cham, pp 141–154 https://doi.org/10.1007/978-3-319-04135-3_9 **(Law, Governance and Technology Series)**

Solove DJ (2005) A taxonomy of privacy. SSRN Scholarly Paper ID 667622. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=667622. Accessed 19 Oct 2018

Stahl BC (2006) Accountability and reflective responsibility in information systems. In: Zielinski C, Duquenoy P, Kimppa K (eds) The information society: emerging landscapes. Springer, US (IFIP International Federation for Information Processing), pp 51–68

Tavani HT (2007) Philosophical theories of privacy: implications for an adequate online privacy policy. Metaphilosophy 38(1):1–22

Tayi GK, Ballou DP (1998) Examining data quality. Commun ACM 41(2):54–57. https://doi.org/10.1145/269012.269021

Tu Y-C (2014) Transparency in software engineering. Thesis. ResearchSpace@Auckland. https://researchspace.auckland.ac.nz/handle/2292/22092. Accessed 19 Oct 2018

Turilli M, Floridi L (2009) The ethics of information transparency. Ethics Inf Technol 11(2):105–112. https://doi.org/10.1007/s10676-009-9187-9

Vallor S (2018) Technology and the virtues: a philosophical guide to a future worth wanting. Reprint edition. Oxford University Press

Van de Poel I (2011) The relation between forward-looking and backward-looking responsibility. In: Vincent NA, Poel I, Hoven J (eds) Moral responsibility: beyond free will and determinism. Springer, Dordrecht, pp 37–52 https://doi.org/10.1007/978-94-007-1878-4_3 **(Library of Ethics and Applied Philosophy)**

Van de Poel I (2013) Translating values into design requirements. In: Michelfelder DP, McCarthy N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht, pp 253–266 https://doi.org/10.1007/978-94-007-7762-0_20 **(Philosophy of Engineering and Technology)**

Van de Poel I (2018, draft) Core values and value conflicts in cybersecurity; beyond privacy versus security. In: Christen M, Loi M, Gordijn B (eds) The ethics of cybersecurity. Springer, Dordrecht

Van de Poel I, Royakkers L (2011) Ethics, technology, and engineering: an introduction, 1st edn. Wiley-Blackwell, Malden

Van den Hoven J, Lokhorst G-J, van de Poel I (2012) Engineering and the problem of moral overload. Sci Eng Ethics 18(1):143–155. https://doi.org/10.1007/s11948-011-9277-z

Van den Hoven MJ (1998) Moral responsibility, public office and information technology. In: Snellen ITM, Donk WBH (eds) Public administration in an information age: a handbook. IOS Press, Amsterdam, pp 97–112

Van der Velden L (2015) Leaky apps and data shots: technologies of leakage and insertion in NSA-surveillance. Surveill Soc 13(2):182–196. https://doi.org/10.24908/ss.v13i2.5315

Van der Voort HG, Klievink AJ, Arnaboldi M, Meijir AJ (2019) Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? Gov Inf Q 36(1):27–38. https://doi.org/10.1016/j.giq.2018.10.011

Van der Voort HG et al (2019) Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? Gov Inf Q 36(1):27–38. https://doi.org/10.1016/j.giq.2018.10.011

Van Dijck J (2014) Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. Surveill Soc 12(2):197–208. https://doi.org/10.24908/ss.v12i2.4776

Walzer M (1983) Spheres of justice: a defense of pluralism and equality, Reprint edn. Basic Books, New York

Warren SD, Brandeis LD (1890) The right to privacy. Harv Law Rev 4(5):193–220. https://doi.org/10.2307/1321160

Weizenbaum J (1977) Computer power and human reason: from judgement to calculation. New edition. W.H.Freeman & Co Ltd., San Francisco

Wenar L (2017) John Rawls. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/rawls/. Accessed 25 Oct 2018

Winston A (2018) Palantir has secretly been using New Orleans to test its predictive policing technology, The Verge. https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd. Accessed 26 Mar 2019