BIG DATA & SOCIETY

Original Research Article

Big Data & Society
July–December 2016: 1–21
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2053951716679679
bds.sagepub.com
SAGE

The ethics of algorithms: Mapping the debate

Brent Daniel Mittelstadt[1], Patrick Allo[1], Mariarosaria Taddeo[1,2], Sandra Wachter[2] and Luciano Floridi[1,2]

Abstract

In information societies, operations, decisions and choices previously left to humans are increasingly delegated to algorithms, which may advise, if not decide, about how data should be interpreted and what actions should be taken as a result. More and more often, algorithms mediate social processes, business transactions, governmental decisions, and how we perceive, understand, and interact among ourselves and with the environment. Gaps between the design and operation of algorithms and our understanding of their ethical implications can have severe consequences affecting individuals as well as groups and whole societies. This paper makes three contributions to clarify the ethical importance of algorithmic mediation. It provides a prescriptive map to organise the debate. It reviews the current discussion of ethical aspects of algorithms. And it assesses the available literature in order to identify areas requiring further work to develop the ethics of algorithms.

Keywords

Algorithms, automation, Big Data, data analytics, data mining, ethics, machine learning

Introduction

In information societies, operations, decisions and choices previously left to humans are increasingly delegated to algorithms, which may advise, if not decide, about how data should be interpreted and what actions should be taken as a result.[1] Examples abound. Profiling and classification algorithms determine how individuals and groups are shaped and managed (Floridi, 2012). Recommendation systems give users directions about when and how to exercise, what to buy, which route to take, and who to contact (Vries, 2010: 81). Data mining algorithms are said to show promise in helping make sense of emerging streams of behavioural data generated by the 'Internet of Things' (Portmess and Tower, 2014: 1). Online service providers continue to mediate how information is accessed with personalisation and filtering algorithms (Newell and Marabelli, 2015; Taddeo and Floridi, 2015). Machine learning algorithms automatically identify misleading, biased or inaccurate knowledge at the point of creation (e.g. Wikipedia's Objective Revision Evaluation Service). As these examples suggest, how we perceive and understand our environments and interact with them and each other is increasingly mediated by algorithms.

Algorithms are inescapably value-laden (Brey and Soraker, 2009; Wiener, 1988). Operational parameters are specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others (cf. Friedman and Nissenbaum, 1996; Johnson, 2006; Kraemer et al., 2011; Nakamura, 2013). At the same time, operation within accepted parameters does not guarantee ethically acceptable behaviour. This is shown, for example, by profiling algorithms that inadvertently discriminate against marginalised populations (Barocas and Selbst,

[1]Oxford Internet Institute, University of Oxford, Oxford, UK
[2]Alan Turing Institute, British Library, London, UK

Corresponding author:
Brent Daniel Mittelstadt, Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK.
Email: brent.mittelstadt@oii.ox.ac.uk

2015; Birrer, 2005), as seen in delivery of online advertisements according to perceived ethnicity (Sweeney, 2013).

Determining the potential and actual ethical impact of an algorithm is difficult for many reasons. Identifying the influence of human subjectivity in algorithm design and configuration often requires investigation of long-term, multi-user development processes. Even with sufficient resources, problems and underlying values will often not be apparent until a problematic use case arises. Learning algorithms, often quoted as the 'future' of algorithms and analytics (Tutt, 2016), introduce uncertainty over how and why decisions are made due to their capacity to tweak operational parameters and decision-making rules 'in the wild' (Burrell, 2016). Determining whether a particular problematic decision is merely a one-off 'bug' or evidence of a systemic failure or bias may be impossible (or at least highly difficult) with poorly interpretable and predictable learning algorithms. Such challenges are set to grow, as algorithms increase in complexity and interact with each other's outputs to take decisions (Tutt, 2016). The resulting gap between the design and operation of algorithms and our understanding of their ethical implications can have severe consequences affecting individuals, groups and whole segments of a society.

In this paper, we map the ethical problems prompted by algorithmic decision-making. The paper answers two questions: what kinds of ethical issues are raised by algorithms? And, how do these issues apply to algorithms themselves, as opposed to technologies built upon algorithms? We first propose a conceptual map based on six kinds of concerns that are jointly sufficient for a principled organisation of the field. We argue that the map allows for a more rigorous diagnosis of ethical challenges related to the use of algorithms. We then review the scientific literature discussing ethical aspects of algorithms to assess the utility and accuracy of the proposed map. Seven themes emerged from the literature that demonstrate how the concerns defined in the proposed map arise in practice. Together, the map and review provide a common structure for future discussion of the ethics of algorithms. In the final section of the paper we assess the fit between the proposed map and themes currently raised in the reviewed literature to identify areas of the 'ethics of algorithms' requiring further research. The conceptual framework, review and critical analysis offered in this paper aim to inform future ethical inquiry, development, and governance of algorithms.

Background

To map the ethics of algorithms, we must first define some key terms. 'Algorithm' has an array of meanings across computer science, mathematics and public discourse. As Hill explains, "we see evidence that any procedure or decision process, however ill-defined, can be called an 'algorithm' in the press and in public discourse. We hear, in the news, of 'algorithms' that suggest potential mates for single people and algorithms that detect trends of financial benefit to marketers, with the implication that these algorithms may be right or wrong..." (Hill, 2015: 36). Many scholarly critiques also fail to specify technical categories or a formal definition of 'algorithm' (Burrell, 2016; Kitchin, 2016). In both cases the term is used not in reference to the algorithm as a mathematical construct, but rather the implementation and interaction of one or more algorithms in a particular program, software or information system. Any attempt to map an 'ethics of algorithms' must address this conflation between formal definitions and popular usage of 'algorithm'.

Here, we follow Hill's (2015: 47) formal definition of an algorithm as a *mathematical construct* with "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions." However, our investigation will not be limited to algorithms as mathematical constructs. As suggested by the inclusion of 'purpose' and 'provisions' in Hill's definition, algorithms must be implemented and executed to take action and have effects. The popular usage of the term becomes relevant here. References to algorithms in public discourse do not normally address algorithms as mathematical constructs, but rather particular implementations. Lay usage of 'algorithm' also includes *implementation* of the mathematical construct into a technology, and an application of the technology *configured* for a particular task.[2] A fully configured algorithm will incorporate the abstract mathematical structure that has been implemented into a system for analysis of tasks in a particular analytic domain. Given this clarification, the configuration of an algorithm to a specific task or dataset does not change its underlying mathematical representation or system implementation; it is rather a further tweaking of the algorithm's operation in relation to a specific case or problem.

Accordingly, it makes little sense to consider the ethics of algorithms independent of how they are implemented and executed in computer programs, software and information systems. Our aim here is to map the ethics of algorithms, with 'algorithm' interpreted along public discourse lines. Our map will include ethical issues arising from algorithms as mathematical constructs, implementations (technologies, programs) and configurations (applications).[3] Where discussion focuses on implementations or configurations (i.e. an artefact with an embedded algorithm), we limit our focus to issues relating to the algorithm's work, rather than all issues related to the artefact.

However, as noted by Hill above, a problem with the popular usage of 'algorithm' is that it can describe ''any procedure or decision process,'' resulting in a prohibitively large range of artefacts to account for in a mapping exercise. Public discourse is currently dominated by concerns with a particular class of algorithms that make decisions, e.g. the best action to take in a given situation, the best interpretation of data, and so on. Such algorithms augment or replace analysis and decision-making by humans, often due to the scope or scale of data and rules involved. Without offering a precise definition of the class, the algorithms we are interested in here are those that make generally reliable (but subjective and not necessarily correct) decisions based upon complex rules that challenge or confound human capacities for action and comprehension.[4] In other words, we are interested in algorithms whose actions are difficult for humans to predict or whose decision-making logic is difficult to explain after the fact. Algorithms that automate mundane tasks, for instance in manufacturing, are not our concern.

Decision-making algorithms are used across a variety of domains, from simplistic decision-making models (Levenson and Pettrey, 1994) to complex profiling algorithms (Hildebrandt, 2008). Notable contemporary examples include online software agents used by online service providers to carry out operations on the behalf of users (Kim et al., 2014); online dispute resolution algorithms that replace human decision-makers in dispute mediation (Raymond, 2014; Shackelford and Raymond, 2014); recommendation and filtering systems that compare and group users to provide personalised content (Barnet, 2009); clinical decision support systems (CDSS) that recommend diagnoses and treatments to physicians (Diamond et al., 1987; Mazoué, 1990); and predictive policing systems that predict criminal activity hotspots.

The discipline of data analytics is a standout example, defined here as the practice of using algorithms to make sense of streams of data. Analytics informs immediate responses to the needs and preferences of the users of a system, as well as longer term strategic planning and development by a platform or service provider (Grindrod, 2014). Analytics identifies relationships and small patterns across vast and distributed datasets (Floridi, 2012). New types of enquiry are enabled, including behavioural research on 'scraped' data (e.g. Lomborg and Bechmann, 2014: 256); tracking of fine-grained behaviours and preferences (e.g. sexual orientation or political opinions; (Mahajan et al., 2012); and prediction of future behaviour (as used in predictive policing or credit, insurance and employment screening; Zarsky, 2016). *Actionable insights* (more on this later) are sought rather than causal relationships (Grindrod, 2014; Hildebrandt, 2011; Johnson, 2013).

Analytics demonstrates how algorithms can challenge human decision-making and comprehension even for tasks previously performed by humans. In making a decision (for instance, which risk class a purchaser of insurance belongs to), analytics algorithms work with high-dimension data to determine which features are relevant to a given decision. The number of features considered in any such classification task can run into the tens of thousands. This type of task is thus a replication of work previously undertaken by human workers (i.e. risk stratification), but involving a qualitatively different decision-making logic applied to greater inputs.

Algorithms are, however, ethically challenging not only because of the scale of analysis and complexity of decision-making. The uncertainty and opacity of the work being done by algorithms and its impact is also increasingly problematic. Algorithms have traditionally required decision-making rules and weights to be individually defined and programmed 'by hand'. While still true in many cases (Google's PageRank algorithm is a standout example), algorithms increasingly rely on learning capacities (Tutt, 2016).

Machine learning is ''any methodology and set of techniques that can employ data to come up with novel patterns and knowledge, and generate models that can be used for effective predictions about the data'' (Van Otterlo, 2013). Machine learning is defined by the capacity to define or modify decision-making rules autonomously. A machine learning algorithm applied to classification tasks, for example, typically consists of two components, a *learner* which produces a *classifier*, with the intention to develop classes that can generalise beyond the training data (Domingos, 2012). The algorithm's work involves placing new inputs into a model or classification structure. Image recognition technologies, for example, can decide what types of objects appear in a picture. The algorithm 'learns' by defining rules to determine how new inputs will be classified. The model can be taught to the algorithm via hand labelled inputs (supervised learning); in other cases the algorithm itself defines best-fit models to make sense of a set of inputs (unsupervised learning)[5] (Schermer, 2011; Van Otterlo, 2013). In both cases, the algorithm defines decision-making rules to handle new inputs. Critically, the human operator does not need to understand the rationale of decision-making rules produced by the algorithm (Matthias, 2004: 179).

As this explanation suggests, learning capacities grant algorithms some degree of autonomy. The impact of this autonomy must remain uncertain to some degree. As a result, tasks performed by machine learning are difficult to predict beforehand (how a new input will be handled) or explain afterwards (how a particular decision was made). Uncertainty can thus inhibit the identification and redress of ethical challenges in the design and operation of algorithms.

Map of the ethics of algorithms

Using the key terms defined in the previous section, we propose a conceptual map (Figure 1) based on six types of concerns that are jointly sufficient for a principled organisation of the field, and conjecture that it allows for a more rigorous diagnosis of ethical challenges related to the use of algorithms. The map is not proposed from a particular theoretical or methodological approach to ethics, but rather is intended as a prescriptive framework of types of issues arising from algorithms owing to three aspects of how algorithms operate. The map takes into account that the algorithms this paper is concerned with are used to (1) turn data into evidence for *a given outcome* (henceforth conclusion), and that this outcome is then used to (2) trigger *and* motivate an action that (on its own, or when combined with other actions) may not be ethically neutral. This work is performed in ways that are complex and (semi-)autonomous, which (3) complicates apportionment of responsibility for effects of actions driven by algorithms. The map is thus not intended as a tool to help solve ethical dilemmas arising from problematic actions driven by algorithms, but rather is posed as an organising structure based on how algorithms operate that can structure future discussion of ethical issues. This leads us to posit three epistemic, and two normative kinds of ethical concerns arising from the use of algorithms, based on how algorithms process data to produce evidence and motivate actions. These concerns are associated with potential failures that may involve multiple actors, and therefore complicate the question of who should be held responsible and/or accountable for such failures. Such difficulties motivate the addition of *traceability* as a final, overarching, concern.

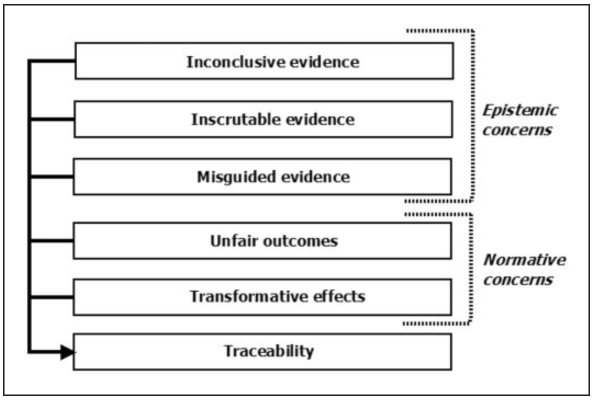

Figure 1. Six types of ethical concerns raised by algorithms.

Inconclusive evidence

When algorithms draw conclusions from the data they process using inferential statistics and/or machine learning techniques, they produce probable[6] yet inevitably uncertain knowledge. Statistical learning theory (James et al., 2013) and computational learning theory (Valiant, 1984) are both concerned with the characterisation and quantification of this uncertainty. In addition to this, and as often indicated, statistical methods can help identify significant correlations, but these are rarely considered to be sufficient to posit the existence of a causal connection (Illari and Russo, 2014: Chapter 8), and thus may be insufficient to motivate action on the basis of knowledge of such a connection. The term *actionable insight* we mentioned earlier can be seen as an explicit recognition of these epistemic limitations.

Algorithms are typically deployed in contexts where more reliable techniques are either not available or too costly to implement, and are thus rarely meant to be infallible. Recognising this limitation is important, but should be complemented with an assessment of how the risk of being wrong affects one's epistemic responsibilities (Miller and Record, 2013): for instance, by weakening the justification one has for a conclusion beyond what would be deemed acceptable to justify action in the context at hand.

Inscrutable evidence

When data are used as (or processed to produce) evidence for a conclusion, it is reasonable to expect that the connection between the data and the conclusion should be accessible (i.e. intelligible as well as open to scrutiny and perhaps even critique).[7] When the connection is not obvious, this expectation can be satisfied by better access as well as by additional explanations. Given how algorithms operate, these requirements are not automatically satisfied. A lack of knowledge regarding the data being used (e.g. relating to their scope, provenance and quality), but more importantly also the inherent difficulty in the interpretation of how each of the many data-points used by a machine-learning algorithm contribute to the conclusion it generates, cause practical as well as principled limitations (Miller and Record, 2013).

Misguided evidence

Algorithms process data and are therefore subject to a limitation shared by all types of data-processing, namely that the output can never exceed the input.

While Shannon's mathematical theory of communication (Shannon and Weaver, 1998), and especially some of his information-inequalities, give a formally precise account of this fact, the informal 'garbage in, garbage out' principle clearly illustrates what is at stake here, namely that conclusions can only be as reliable (but also as neutral) as the data they are based on. Evaluations of the neutrality of the process, and by connection whether the evidence produced is misguided, are of course observer-dependent.

Unfair outcomes

The three epistemic concerns detailed thus far address the quality of *evidence* produced by an algorithm that motivates a particular action. However, ethical evaluation of algorithms can also focus solely on the *action* itself. Actions driven by algorithms can be assessed according to numerous ethical criteria and principles, which we generically refer to here as the observer-dependent 'fairness' of the action and its effects. An action can be found discriminatory, for example, solely from its effect on a protected class of people, even if made on the basis of conclusive, scrutable and well-founded evidence.

Transformative effects

The ethical challenges posed by the spreading use of algorithms cannot always be retraced to clear cases of epistemic or ethical failures, for some of the effects of the reliance on algorithmic data-processing and (semi-) autonomous decision-making can be questionable and yet appear ethically neutral because they do not seem to cause any obvious harm. This is because algorithms can affect how we conceptualise the world, and modify its social and political organisation (cf. Floridi, 2014). Algorithmic activities, like profiling, reontologise the world by understanding and conceptualising it in new, unexpected ways, and triggering and motivating actions based on the insights it generates.

Traceability

Algorithms are software-artefacts used in data-processing, and as such inherit the ethical challenges associated with the design and availability of new technologies and those associated with the manipulation of large volumes of personal and other data. This implies that harm caused by algorithmic activity is hard to debug (i.e. to detect the harm and find its cause), but also that it is rarely straightforward to identify who should be held responsible for the harm caused.[8] When a problem is identified addressing any or all of the five preceding kinds, ethical assessment requires both the cause and responsibility for the harm to be traced.

Thanks to this map (Figure 1), we are now able to distinguish epistemological, strictly ethical and traceability types in descriptions of ethical problems with algorithms. The map is thus intended as a tool to organise a widely dispersed academic discourse addressing a diversity of technologies united by their reliance on algorithms. To assess the utility of the map, and to observe how each of these kinds of concerns manifests in ethical problems already observed in algorithms, a systematic review of academic literature was carried out.[9] The following sections (4 to 10) describe how ethical issues and concepts are treated in the literature explicitly discussing the ethical aspects of algorithms.

Inconclusive evidence leading to unjustified actions

Much algorithmic decision-making and data mining relies on inductive knowledge and correlations identified within a dataset. Causality is not established prior to acting upon the evidence produced by the algorithm. The search for causal links is difficult, as correlations established in large, proprietary datasets are frequently not reproducible or falsifiable (cf. Ioannidis, 2005; Lazer et al., 2014). Despite this, correlations based on a sufficient volume of data are increasingly seen as sufficiently credible to direct action without first establishing causality (Hildebrandt, 2011; Hildebrandt and Koops, 2010; Mayer-Schönberger and Cukier, 2013; Zarsky, 2016). In this sense data mining and profiling algorithms often need only establish a sufficiently reliable evidence base to drive action, referred to here as actionable insights.

Acting on correlations can be doubly problematic.[10] Spurious correlations may be discovered rather than genuine causal knowledge. In predictive analytics correlations are doubly uncertain (Ananny, 2016). Even if strong correlations or causal knowledge are found, this knowledge may only concern populations while actions are directed towards individuals (Illari and Russo, 2014). As Ananny (2016: 103) explains, ''algorithmic categories ... signal certainty, discourage alternative explorations, and create coherence among disparate objects,'' all of which contribute to individuals being described (possibly inaccurately) via simplified models or classes (Barocas, 2014). Finally, even if both actions and knowledge are at the population-level, our actions may spill over into the individual level. For example, this happens when an insurance premium is set for a sub-population, and hence has to be paid by each member. Actions taken on the basis of inductive correlations have real impact on human interests independent of their validity.

Inscrutable evidence leading to opacity

The scrutability of evidence, evaluated in terms of the transparency or opacity of algorithms, proved a major concern in the reviewed literature. Transparency is generally desired because algorithms that are poorly predictable or explainable are difficult to control, monitor and correct (Tutt, 2016). As many critics have observed (Crawford, 2016; Neyland, 2016; Raymond, 2014), transparency is often naïvely treated as a panacea for ethical issues arising from new technologies. Transparency is generally defined with respect to ''the availability of information, the conditions of accessibility and how the information ... may pragmatically or epistemically support the user's decision-making process'' (Turilli and Floridi, 2009: 106). The debate on this topic is not new. The literature in information and computer ethics, for example started to focus on it at the beginning of the 21st century, when issues concerning algorithmic information filtering by search engines arose.[11]

The primary components of transparency are *accessibility* and *comprehensibility* of information. Information about the functionality of algorithms is often intentionally poorly accessible. Proprietary algorithms are kept secret for the sake of competitive advantage (Glenn and Monteith, 2014; Kitchin, 2016; Stark and Fins, 2013), national security (Leese, 2014), or privacy. Transparency can thus run counter to other ethical ideals, in particular the privacy of data subjects and autonomy of organisations.

Granka (2010) notes a power struggle between data subjects' interests in transparency and data processors' commercial viability. Disclosing the structure of these algorithms would facilitate ill-intentioned manipulations of search results (or 'gaming the system'), while not bringing any advantage to the average non-tech-savvy user (Granka, 2010; Zarsky, 2016). The commercial viability of data processors in many industries (e.g. credit reporting, high frequency trading) may be threatened by transparency. However, data subjects retain an interest in understanding how information about them is created and influences decisions taken in data-driven practices. This struggle is marked by information asymmetry and an ''imbalance in knowledge and decision-making power'' favouring data processors (Tene and Polonetsky, 2013a: 252).

Besides being accessible, information must be comprehensible to be considered transparent (Turilli and Floridi, 2009). Efforts to make algorithms transparent face a significant challenge to render complex decision-making processes both accessible and comprehensible. The longstanding problem of interpretability in machine learning algorithms indicates the challenge of opacity in algorithms (Burrell, 2016; Hildebrandt, 2011; Leese, 2014; Tutt, 2016). Machine learning is adept at creating and modifying rules to classify or cluster large datasets. The algorithm modifies its behavioural structure during operation (Markowetz et al., 2014). This alteration of how the algorithm classifies new inputs is how it learns (Burrell, 2016: 5). Training produces a structure (e.g. classes, clusters, ranks, weights, etc.) to classify new inputs or predict unknown variables. Once trained, new data can be processed and categorised automatically without operator intervention (Leese, 2014). The rationale of the algorithm is obscured, lending to the portrayal of machine learning algorithms as 'black boxes'.

Burrell (2016) and Schermer (2011) argue that the opacity of machine learning algorithms inhibits oversight. Algorithms ''are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs'' (Burrell, 2016: 1). Both the inputs (data about humans) and outputs (classifications) can be unknown and unknowable. Opacity in machine learning algorithms is a product of the high-dimensionality of data, complex code and changeable decision-making logic (Burrell, 2016). Matthias (2004: 179) suggests that machine learning can produce outputs for which ''the human trainer himself is unable to provide an algorithmic representation.'' Algorithms can only be considered explainable to the degree that a human can articulate the trained model or rationale of a particular decision, for instance by explaining the (quantified) influence of particular inputs or attributes (Datta et al., 2016). Meaningful oversight and human intervention in algorithmic decision-making ''is impossible when the machine has an informational advantage over the operator ... [or] when the machine cannot be controlled by a human in real-time due to its processing speed and the multitude of operational variables'' (Matthias, 2004: 182–183). This is, one again, the black box problem. However, a distinction should be drawn, between the technical infeasibility of oversight and practical barriers caused, for instance, by a lack of expertise, access, or resources.

Beyond machine learning, algorithms with 'handwritten' decision-making rules can still be highly complex and practically inscrutable from a lay data subject's perspective (Kitchin, 2016). Algorithmic decision-making structures containing ''hundreds of rules are very hard to inspect visually, especially when their predictions are combined probabilistically in complex ways'' (Van Otterlo, 2013). Further, algorithms are often developed by large teams of engineers over time, with a holistic understanding of the development process and its embedded values, biases and interdependencies rendered infeasible (Sandvig et al., 2014). In both respects, algorithmic processing contrasts with traditional decision-making, where human decision-makers can in principle articulate their rationale when queried, limited only by their desire and capacity to give an explanation, and the questioner's capacity to understand it. The rationale of an algorithm can in contrast be incomprehensible to humans, rendering the legitimacy of decisions difficult to challenge.

Under these conditions, decision-making is poorly transparent. Rubel and Jones (2014) argue that the failure to render the processing logic comprehensible to data subject's disrespects their agency (we shall return to this point in Section 8). Meaningful consent to data-processing is not possible when opacity precludes risk assessment (Schermer, 2011). Releasing information about the algorithm's decision-making logic in a simplified format can help (Datta et al., 2016; Tene and Polonetsky, 2013a). However, complex decision-making structures can quickly exceed the human and organisational resources available for oversight (Kitchin, 2016). As a result, lay data subjects may lose trust in both algorithms and data processors (Cohen et al., 2014; Rubel and Jones, 2014; Shackelford and Raymond, 2014).[12]

Even if data processors and controllers disclose operational information, the net benefit for society is uncertain. A lack of public engagement with existing transparency mechanisms reflects this uncertainty, seen for example in credit scoring (Zarsky, 2016). Transparency disclosures may prove more impactful if tailored towards trained third parties or regulators representing public interest as opposed to data subjects themselves (Tutt, 2016; Zarsky, 2013).

Transparency disclosures by data processors and controllers may prove crucial in the future to maintain a trusting relationship with data subjects (Cohen et al., 2014; Rubel and Jones, 2014; Shackelford and Raymond, 2014). Trust implies the trustor's (the agent who trusts) expectations for the trustee (the agent who is trusted) to perform a task (Taddeo, 2010), and acceptance of the risk that the trustee will betray these expectations (Wiegel and Berg, 2009). Trust in data processors can, for instance, alleviate concerns with opaque personal data-processing (Mazoué, 1990). However, trust can also exist among artificial agents exclusively, seen for instance in the agents of a distributed system working cooperatively to achieve a given goal (Grodzinsky et al., 2010; Simon, 2010; Taddeo, 2010). Furthermore, algorithms can be perceived as trustworthy independently of (or perhaps even despite) any trust placed in the data processor; yet the question of when this may be appropriate remains open.[13]

Misguided evidence leading to bias

The automation of human decision-making is often justified by an alleged lack of bias in algorithms (Bozdag, 2013; Naik and Bhide, 2014). This belief is unsustainable, as shown by prior work demonstrating the normativity of information technologies in general and algorithm development in particular[14] (e.g. Bozdag, 2013; Friedman and Nissenbaum, 1996; Kraemer et al., 2011; Macnish, 2012; Newell and Marabelli, 2015: 6; Tene and Polonetsky, 2013b). Much of the reviewed literature addresses how bias manifests in algorithms and the evidence they produce.

Algorithms inevitably make biased decisions. An algorithm's design and functionality reflects the values of its designer and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option. Development is not a neutral, linear path; there is no objectively correct choice at any given stage of development, but many possible choices (Johnson, 2006). As a result, ''the values of the author [of an algorithm], wittingly or not, are frozen into the code, effectively institutionalising those values'' (Macnish, 2012: 158). It is difficult to detect latent bias in algorithms and the models they produce when encountered in isolation of the algorithm's development history (Friedman and Nissenbaum, 1996; Hildebrandt, 2011; Morek, 2006).

Friedman and Nissenbaum (1996) argue that bias can arise from (1) pre-existing social values found in the ''social institutions, practices and attitudes'' from which the technology emerges, (2) technical constraints and (3) emergent aspects of a context of use. Social biases can be embedded in system design purposefully by individual designers, seen for instance in manual adjustments to search engine indexes and ranking criteria (Goldman, 2006). Social bias can also be unintentional, a subtle reflection of broader cultural or organisational values. For example, machine learning algorithms trained from human-tagged data inadvertently learn to reflect biases of the taggers (Diakopoulos, 2015).

Technical bias arises from technological constraints, errors or design decisions, which favour particular groups without an underlying driving value (Friedman and Nissenbaum, 1996). Examples include when an alphabetical listing of airline companies leads to increase business for those earlier in the alphabet, or an error in the design of a random number generator that causes particular numbers to be favoured. Errors can similarly manifest in the datasets processed by algorithms. Flaws in the data are inadvertently adopted by the algorithm and hidden in outputs and models produced (Barocas and Selbst, 2015; Romei and Ruggieri, 2014).

Emergent bias is linked with advances in knowledge or changes to the system's (intended) users and stakeholders (Friedman and Nissenbaum, 1996). For example, CDSS are unavoidably biased towards treatments included in their decision architecture. Although emergent bias is linked to the user, it can emerge unexpectedly from decisional rules developed by the algorithm, rather than any 'hand-written' decision-making structure (Hajian and Domingo-Ferrer, 2013; Kamiran and Calders, 2010). Human monitoring may prevent some biases from entering algorithmic decision-making in these cases (Raymond, 2014).

The outputs of algorithms also require interpretation (i.e. what one should do based on what the algorithm indicates); for behavioural data, 'objective' correlations can come to reflect the interpreter's "unconscious motivations, particular emotions, deliberate choices, socio-economic determinations, geographic or demographic influences" (Hildebrandt, 2011: 376). Explaining the correlation in any of these terms requires additional justification – meaning is not self-evident in statistical models. Different metrics "make visible aspects of individuals and groups that are not otherwise perceptible" (Lupton, 2014: 859). It thus cannot be assumed that an observer's interpretation will correctly reflect the perception of the actor rather than the biases of the interpreter.

Unfair outcomes leading to discrimination

Much of the reviewed literature also addresses how discrimination results from biased evidence and decision-making.[15] Profiling by algorithms, broadly defined "as the construction or inference of patterns by means of data mining and . . . the application of the ensuing profiles to people whose data match with them" (Hildebrandt and Koops, 2010: 431), is frequently cited as a source of discrimination. Profiling algorithms identify correlations and make predictions about behaviour at a group-level, albeit with groups (or profiles) that are constantly changing and re-defined by the algorithm (Zarsky, 2013). Whether dynamic or static, the individual is comprehended based on connections with others identified by the algorithm, rather than actual behaviour (Newell and Marabelli, 2015: 5). Individuals' choices are structured according to information about the group (Danna and Gandy, 2002: 382). Profiling can inadvertently create an evidence-base that leads to discrimination (Vries, 2010).

For the affected parties, data-driven discriminatory treatment is unlikely to be more palatable than discrimination fuelled by prejudices or anecdotal evidence. This much is implicit in Schermer's (2011) argument that discriminatory treatment is not ethically problematic in itself; rather, it is the effects of the treatment that determine its ethical acceptability. However, Schermer muddles bias and discrimination into a single concept. What he terms discrimination can be described instead as mere bias, or the consistent and repeated expression of a particular preference, belief or value in decision-making (Friedman and Nissenbaum, 1996). In contrast, what he describes as problematic effects of discriminatory treatment can be defined as discrimination *tout court*. So bias is a dimension of the decision-making itself, whereas discrimination describes the effects of a decision, in terms of adverse disproportionate impact resulting from algorithmic decision-making. Barocas and Selbst (2015) show that precisely this definition guides 'disparate impact detection', an enforcement mechanism for American anti-discrimination law in areas such as social housing and employment. They suggest that disparate impact detection provides a model for the detection of bias and discrimination in algorithmic decision-making which is sensitive to differential privacy.

It may be possible to direct algorithms not to consider sensitive attributes that contribute to discrimination (Barocas and Selbst, 2015), such as gender or ethnicity (Calders et al., 2009; Kamiran and Calders, 2010; Schermer, 2011), based upon the emergence of discrimination in a particular context. However, proxies for protected attributes are not easy to predict or detect (Romei and Ruggieri, 2014; Zarsky, 2016), particularly when algorithms access linked datasets (Barocas and Selbst, 2015). Profiles constructed from neutral characteristics such as postal code may inadvertently overlap with other profiles related to ethnicity, gender, sexual preference, and so on (Macnish, 2012; Schermer, 2011).

Efforts are underway to avoid such 'redlining' by sensitive attributes and proxies. Romei and Ruggieri (2014) observe four overlapping strategies for discrimination prevention in analytics: (1) controlled distortion of training data; (2) integration of anti-discrimination criteria into the classifier algorithm; (3) post-processing of classification models; (4) modification of predictions and decisions to maintain a fair proportion of effects between protected and unprotected groups. These strategies are seen in the development of privacy-preserving, fairness- and discrimination-aware data mining (Dwork et al., 2011; Kamishima et al., 2012). Fairness-aware data mining takes the broadest aim, as it gives attention not only to discrimination but fairness, neutrality, and independence as well (Kamishima et al., 2012). Various metrics of fairness are possible based on statistical parity, differential privacy and other relations between data subjects in classification tasks (Dwork et al., 2011; Romei and Ruggieri, 2014).

The related practice of personalisation is also frequently discussed. Personalisation can segment a population so that only some segments are worthy of receiving some opportunities or information, re-enforcing existing social (dis)advantages. Questions of the fairness and equitability of such practices are often raised (e.g. Cohen et al., 2014; Danna and Gandy, 2002; Rubel and Jones, 2014). Personalised pricing, for example, can be ‘‘an invitation to leave quietly’’ issued to data subjects deemed to lack value or the capacity to pay.[16]

Reasons to consider discriminatory effects as *adverse* and thus ethically problematically are diverse. Discriminatory analytics can contribute to self-fulfilling prophecies and stigmatisation in targeted groups, undermining their autonomy and participation in society (Barocas, 2014; Leese, 2014; Macnish, 2012). Personalisation through non-distributive profiling, seen for example in personalised pricing in insurance premiums (Hildebrandt and Koops, 2010; Van Wel and Royakkers, 2004), can be discriminatory by violating both ethical and legal principles of equal or fair treatment of individuals (Newell and Marabelli, 2015). Further, as described above the capacity of individuals to investigate the personal relevance of factors used in decision-making is inhibited by opacity and automation (Zarsky, 2016).

Transformative effects leading to challenges for autonomy

Value-laden decisions made by algorithms can also pose a threat to the autonomy of data subjects. The reviewed literature in particular connects personalisation algorithms to these threats. Personalisation can be defined as the construction of choice architectures which are not the same across a sample (Tene and Polonetsky, 2013a). Similar to explicitly persuasive technologies, algorithms can nudge the behaviour of data subjects and human decision-makers by filtering information (Ananny, 2016). Different content, information, prices, etc. are offered to groups or classes of people within a population according to a particular attribute, e.g. the ability to pay.

Personalisation algorithms tread a fine line between supporting and controlling decisions by filtering which information is presented to the user based upon in-depth understanding of preferences, behaviours, and perhaps vulnerabilities to influence (Bozdag, 2013; Goldman, 2006; Newell and Marabelli, 2015; Zarsky, 2016). Classifications and streams of behavioural data are used to match information to the interests and attributes of data subjects. The subject’s autonomy in decision-making is disrespected when the desired choice reflects third party interests above the individual’s (Applin and Fischer, 2015; Stark and Fins, 2013).

This situation is somewhat paradoxical. In principle, personalisation should improve decision-making by providing the subject with only relevant information when confronted with a potential information overload; however, deciding which information is relevant is inherently subjective. The subject can be pushed to make the ‘‘institutionally preferred action rather than their own preference’’ (Johnson, 2013); online consumers, for example, can be nudged to fit market needs by filtering how products are displayed (Coll, 2013). Lewis and Westlund (2015: 14) suggest that personalisation algorithms need to be taught to ‘act ethically’ to strike a balance between coercing and supporting users’ decisional autonomy.

Personalisation algorithms reduce the diversity of information users encounter by excluding content deemed irrelevant or contradictory to the user’s beliefs (Barnet, 2009; Pariser, 2011). Information diversity can thus be considered an enabling condition for autonomy (van den Hoven and Rooksby, 2008). Filtering algorithms that create ‘echo chambers’ devoid of contradictory information may impede decisional autonomy (Newell and Marabelli, 2015). Algorithms may be unable to replicate the ‘‘spontaneous discovery of new things, ideas and options’’ which appear as anomalies against a subject’s profiled interests (Raymond, 2014). With near ubiquitous access to information now feasible in the internet age, issues of access concern whether the ‘right’ information can be accessed, rather than any information at all. Control over personalisation and filtering mechanisms can enhance user autonomy, but potentially at the cost of information diversity (Bozdag, 2013). Personalisation algorithms, and the underlying practice of analytics, can thus both enhance and undermine the agency of data subjects.

Transformative effects leading to challenges for informational privacy

Algorithms are also driving a transformation of notions of privacy. Responses to discrimination, de-individualisation and the threats of opaque decision-making for data subjects’ agency often appeal to informational privacy (Schermer, 2011), or the right of data subjects to ‘‘shield personal data from third parties.’’ Informational privacy concerns the capacity of an individual to control information about herself (Van Wel and Royakkers, 2004), and the effort required by third parties to obtain this information.

A right to identity derived from informational privacy interests suggests that opaque or secretive profiling is problematic.[17] Opaque decision-making by algorithms (see 'Inconclusive evidence leading to unjustified actions' section) inhibits oversight and informed decision-making concerning data sharing (Kim et al., 2014). Data subjects cannot define privacy norms to govern all types of data generically because their value or insightfulness is only established through processing (Hildebrandt, 2011; Van Wel and Royakkers, 2004).

Beyond opacity, privacy protections based upon identifiability are poorly suited to limit external management of identity via analytics. Identity is increasingly influenced by knowledge produced through analytics that makes sense of growing streams of behavioural data. The 'identifiable individual' is not necessarily a part of these processes. Schermer (2011) argues that informational privacy is an inadequate conceptual framework because profiling makes the identifiability of data subjects irrelevant.

Profiling seeks to assemble individuals into meaningful groups, for which identity is irrelevant (Floridi, 2012; Hildebrandt, 2011; Leese, 2014). Van Wel and Royakkers (2004: 133) argue that external identity construction by algorithms is a type of de-individualisation, or a "tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merit." Individuals need never be identified when the profile is assembled to be affected by the knowledge and actions derived from it (Louch et al., 2010: 4). The individual's informational identity (Floridi, 2011) is breached by meaning generated by algorithms that link the subject to others within a dataset (Vries, 2010).

Current regulatory protections similarly struggle to address the informational privacy risks of analytics. 'Personal data' is defined in European data protection law as data describing an *identifiable* person; anonymised and aggregated data are not considered personal data (European Commission, 2012). Privacy preserving data mining techniques which do not require access to individual and identifiable records may mitigate these risks (Agrawal and Srikant, 2000; Fule and Roddick, 2004). Others suggest a mechanism to 'opt-out' of profiling for a particular purpose or context would help protect data subjects' privacy interests (Hildebrandt, 2011; Rubel and Jones, 2014). A lack of recourse mechanisms for data subjects to question the validity of algorithmic decisions further exacerbates the challenges of controlling identity and data about oneself (Schermer, 2011). In response, Hildebrandt and Koops (2010) call for 'smart transparency' by designing the socio-technical infrastructures responsible for profiling in a way that allows individuals to anticipate and respond to how they are profiled.

Traceability leading to moral responsibility

When a technology fails, blame and sanctions must be apportioned. One or more of the technology's designer (or developer), manufacturer or user are typically held accountable. Designers and users of algorithms are typically blamed when problems arise (Kraemer et al., 2011: 251). Blame can only be justifiably attributed when the actor has some degree of control (Matthias, 2004) and intentionality in carrying out the action.

Traditionally, computer programmers have had "control of the behaviour of the machine in every detail" insofar as they can explain its design and function to a third party (Matthias, 2004). This traditional conception of responsibility in software design assumes the programmer can reflect on the technology's likely effects and potential for malfunctioning (Floridi et al., 2014), and make design choices to choose the most desirable outcomes according to the functional specification (Matthias, 2004). With that said, programmers may only retain control in principle due to the complexity and volume of code (Sandvig et al., 2014), and the use of external libraries often treated by the programmer as 'black boxes' (cf. Note 7).

Superficially, the traditional, linear conception of responsibility is suitable to non-learning algorithms. When decision-making rules are 'hand-written', their authors retain responsibility (Bozdag, 2013). Decision-making rules determine the relative weight given to the variables or dimensions of the data considered by the algorithm. A popular example is Facebook's EdgeRank personalisation algorithm, which prioritises content based on date of publication, frequency of interaction between author and reader, media type, and other dimensions. Altering the relative importance of each factor changes the relationships users are encouraged to maintain. The party that sets confidence intervals for an algorithm's decision-making structure shares responsibility for the effects of the resultant false positives, false negatives and spurious correlations (Birrer, 2005; Johnson, 2013; Kraemer et al., 2011). Fule and Roddick (2004: 159) suggest operators also have a responsibility to monitor for ethical impacts of decision-making by algorithms because "the sensitivity of a rule may not be apparent to the miner ... the ability to harm or to cause offense can often be inadvertent." Schermer (2011) similarly suggests that data processors should actively searching for errors and bias in their algorithms and models. Human oversight of complex systems as an accountability mechanism may, however, be impossible due to the challenges for transparency already mentioned (see 'Inscrutable evidence leading to opacity' section). Furthermore, humans kept 'in the loop' of automated decision-making may be poorly equipped to identify problems and take corrective actions (Elish, 2016).

Particular challenges arise for algorithms with learning capacities, which defy the traditional conception of designer responsibility. The model requires the system to be well-defined, comprehensible and predictable; complex and fluid systems (i.e. one with countless decision-making rules and lines of code) inhibit holistic oversight of decision-making pathways and dependencies. Machine learning algorithms are particularly challenging in this respect (Burrell, 2016; Matthias, 2004; Zarsky, 2016), seen for instance in genetic algorithms that program themselves. The traditional model of responsibility fails because ''nobody has enough control over the machine's actions to be able to assume the responsibility for them'' (Matthias, 2004: 177).

Allen et al. (2006: 14) concur in discussing the need for 'machine ethics': ''the modular design of systems can mean that no single person or group can fully grasp the manner in which the system will interact or respond to a complex flow of new inputs.'' From traditional, linear programming through to autonomous algorithms, behavioural control is gradually transferred from the programmer to the algorithm and its operating environment (Matthias, 2004: 182). The gap between the designer's control and algorithm's behaviour creates an *accountability gap* (Cardona, 2008) wherein blame can potentially be assigned to several moral agents simultaneously.

Related segments of the literature address the 'ethics of automation', or the acceptability of replacing or augmenting human decision-making with algorithms (Naik and Bhide, 2014). Morek (2006) finds it problematic to assume that algorithms can replace skilled professionals in a like-for-like manner. Professionals have implicit knowledge and subtle skills (cf. Coeckelbergh, 2013; MacIntyre, 2007) that are difficult to make explicit and perhaps impossible to make computable (Morek, 2006). When algorithmic and human decision-makers work in tandem, norms are required to prescribe when and how human intervention is required, particularly in cases like high-frequency trading where real-time intervention is impossible before harms occur (Davis et al., 2013; Raymond, 2014).

Algorithms that make decisions can be considered blameworthy agents (Floridi and Sanders, 2004a; Wiltshire, 2015). The moral standing and capacity for ethical decision-making of algorithms remains a standout question in machine ethics (e.g. Allen et al., 2006; Anderson, 2008; Floridi and Sanders, 2004a). Ethical decisions require agents to evaluate the desirability of different courses of actions which present conflicts between the interests of involved parties (Allen et al., 2006; Wiltshire, 2015).

For some, learning algorithms should be considered moral agents with some degree of moral responsibility. Requirements for moral agency may differ between humans and algorithms; Floridi and Sanders (2004b) and Sullins (2006) argue, for instance, that 'machine agency' requires significant autonomy, interactive behaviour, and a role with causal accountability, to be distinguished from moral responsibility, which requires intentionality. As suggested above, moral agency and accountability are linked. Assigning moral agency to artificial agents can allow human stakeholders to shift blame to algorithms (Crnkovic and Çürüklü, 2011). Denying agency to artificial agents makes designers responsible for the unethical behaviour of their semi-autonomous creations; bad consequences reflect bad design (Anderson and Anderson, 2014; Kraemer et al., 2011; Turilli, 2007). Neither extreme is entirely satisfactory due to the complexity of oversight and the volatility of decision-making structures.

Beyond the nature of moral agency in machines, work in machine ethics also investigates how best to design moral reasoning and behaviours into autonomous algorithms as artificial moral and ethical agents[18] (Anderson and Anderson, 2007; Crnkovic and Çürüklü, 2011; Sullins, 2006; Wiegel and Berg, 2009). Research into this question remains highly relevant because algorithms can be required to make real-time decisions involving ''difficult trade-offs ... which may include difficult ethical considerations'' without an operator (Wiegel and Berg, 2009: 234).

Automation of decision-making creates problems of ethical consistency between humans and algorithms. Turilli (2007) argues algorithms should be constrained ''by the same set of ethical principles'' as the former human worker to ensure consistency within an organisation's ethical standards. However, ethical principles as used by human decision-makers may prove difficult to define and rendered computable. Virtue ethics is also thought to provide rule sets for algorithmic decision-structures which are easily computable. An ideal model for artificial moral agents based on heroic virtues is suggested by Wiltshire (2015), wherein algorithms are trained to be heroic and thus, moral.[19]

Other approaches do not require ethical principles to serve as pillars of algorithmic decision-making frameworks. Bello and Bringsjord (2012) insist that moral reasoning in algorithms should not be structured around classic ethical principles because it does not reflect how humans actually engage in moral decision-making.

Rather, computational cognitive architectures – which allow machines to 'mind read', or attribute mental states to other agents – are required. Anderson and Anderson (2007) suggest algorithms can be designed to mimic human ethical decision-making modelled on empirical research on how intuitions, principles and reasoning interact. At a minimum this debate reveals that a consensus view does not yet exist for how to practically relocate the social and ethical duties displaced by automation (Shackelford and Raymond, 2014).

Regardless of the design philosophy chosen, Friedman and Nissenbaum (1996) argue that developers have a responsibility to design for diverse contexts ruled by different moral frameworks. Following this, Turilli (2007) proposes collaborative development of ethical requirements for computational systems to ground an operational ethical protocol. Consistency can be confirmed between the protocol (consisting of a decision-making structure) and the designer's or organisation's explicit ethical principles (Turilli and Floridi, 2009).

Points of further research

As the preceding discussion demonstrates, the proposed map (see 'Map of the ethics of algorithms' section) can be used to organise current academic discourse describing ethical concerns with algorithms in a principled way, on purely epistemic and ethical grounds. To borrow a concept from software development, the map can remain perpetually 'in beta'. As new types of ethical concerns with algorithms are identified, or if one of the six described types can be separated into two or more types, the map can be revised. Our intention has been to describe the state of academic discourse around the ethics of algorithms, and to propose an organising tool for future work in the field to bridge linguistic and disciplinary gaps. Our hope is that the map will improve the precision of how ethical concerns with algorithms are described in the future, while also serving as a reminder of the limitations of merely methodological, technical or social solutions to challenges raised by algorithms. As the map indicates, ethical concerns with algorithms are multi-dimensional and thus require multi-dimensional solutions.

While the map provides the bare conceptual structure we need, it still must be populated as deployment and critical studies of algorithms proliferate. The seven themes identified in the preceding sections identify where the 'ethics of algorithms' currently lies on the map. With this in mind, in this section we raise a number of topics not yet receiving substantial attention in the reviewed literature related to the transformative effects and traceability of algorithms. These topics can be considered future directions of travel for the ethics of algorithms as the field expands and matures.

Concerning transformative effects, algorithms change how identity is constructed, managed and protected by privacy and data protection mechanisms (see 'Transformative effects leading to challenges for informational privacy' section). Informational privacy and identifiability are typically closely linked; an individual has privacy insofar as she has control over data and information about her. Insofar as algorithms transform privacy by rendering identifiability less important, a theory of privacy responsive to the reduced importance of identifiability and individuality is required.

Van Wel and Royakkers (2004) urge a re-conceptualisation of personal data, where equivalent privacy protections are afforded to 'group characteristics' when used in place of 'individual characteristics' in generating knowledge about or taking actions towards an individual. Further work is required to describe how privacy operates at group level, absent of identifiability (e.g. Mittelstadt and Floridi, 2016; Taylor et al., 2017). Real world mechanisms to enforce privacy in analytics are also required. One proposal mentioned in the reviewed literature is to develop privacy preserving data mining techniques which do not require access to individual and identifiable records (Agrawal and Srikant, 2000; Fule and Roddick, 2004). Related work is already underway to detect discrimination in data mining (e.g. Barocas, 2014; Calders and Verwer, 2010; Hajian et al., 2012), albeit limited to detection of *illegal* discrimination. Further harm detection mechanisms are necessitated by the capacity of algorithms to disadvantage users in indirect and non-obvious ways that exceed legal definitions of discrimination (Sandvig et al., 2014; Tufekci, 2015). It cannot be assumed that algorithms will discriminate according to characteristics observable or comprehensible to humans.

Concerning traceability, two key challenges for apportioning responsibility for algorithms remain under-researched. First, despite a wealth of literature addressing the moral responsibility and agency of algorithms, insufficient attention has been given to *distributed* responsibility, or responsibility as shared across a network of human and algorithmic actors simultaneously (cf. Simon, 2015). The reviewed literature (see 'Traceability leading to moral responsibility' section) addresses the potential moral agency of algorithms, but does not describe methods and principles for apportioning blame or responsibility across a mixed network of human and algorithmic actors.

Second, substantial trust is already placed in algorithms, in some cases affecting a *de-responsibilisation* of human actors, or a tendency to 'hide behind the computer' and assume *automated* processes are correct by default (Zarsky, 2016: 121). Delegating decision-making to algorithms can shift responsibility away from human decision-makers. Similar effects can be observed in mixed networks of human and information systems as already studied in bureaucracies, characterised by reduced feelings of personal responsibility and the execution of otherwise unjustifiable actions (Arendt, 1971). Algorithms involving stakeholders from multiple disciplines can, for instance, lead to each party assuming the others will shoulder ethical responsibility for the algorithm's actions (Davis et al., 2013). Machine learning adds an additional layer of complexity between designers and actions driven by the algorithm, which may justifiably weaken blame placed upon the former. Additional research is needed to understand the prevalence of these effects in algorithm driven decision-making systems, and to discern how to minimise the inadvertent justification of harmful acts.

A related problem concerns malfunctioning and resilience. The need to apportion responsibility is acutely felt when algorithms malfunction. Unethical algorithms can be thought of as malfunctioning software-artefacts that do not operate as intended. Useful distinctions exists between errors of design (types) and errors of operation (tokens), and between the failure to operate as intended (dysfunction) and the presence of unintended side-effects (misfunction) (Floridi et al., 2014). Misfunctioning is distinguished from mere negative side effects by *avoidability*, or the extent to which comparable extant algorithm types accomplish the intended function without the effects in question. These distinctions clarify ethical aspects of algorithms that are strictly related to their functioning, either in the abstract (for instance when we look at raw performance), or as part of a larger decision-making system, and reveals the multi-faceted interaction between intended and actual behaviour.

Both types of malfunctioning imply distinct responsibilities for algorithm and software developers, users and artefacts. Additional work is required to describe fair apportionment of responsibility for dysfunctioning and misfunctioning across large development teams and complex contexts of use. Further work is also required to specify requirements for resilience to malfunctioning as an ethical ideal in algorithm design. Machine learning in particular raises unique challenges, because achieving the intended or "correct" behaviour does not imply the absence of errors[20] (cf. Burrell, 2016) or harmful actions and feedback loops. Algorithms, particularly those embedded in robotics, can for instance be made safely interruptible insofar as harmful actions can be discouraged without the algorithm being encouraged to deceive human users to avoid further interruptions (Orseau and Armstrong, 2016).

Finally, while a degree of transparency is broadly recognised as a requirement for traceability, how to operationalise transparency remains an open question, particularly for machine learning. Merely rendering the code of an algorithm transparent is insufficient to ensure ethical behaviour. Regulatory or methodological requirements for algorithms to be *explainable* or *interpretable* demonstrate the challenge data controllers now face (Tutt, 2016). One possible path to explainability is algorithmic auditing carried out by data processors (Zarsky, 2016), external regulators (Pasquale, 2015; Tutt, 2016; Zarsky, 2016), or empirical researchers (Kitchin, 2016; Neyland, 2016), using ex post audit studies (Adler et al., 2016; Diakopoulos, 2015; Kitchin, 2016; Romei and Ruggieri, 2014; Sandvig et al., 2014), reflexive ethnographic studies in development and testing (Neyland, 2016), or reporting mechanisms designed into the algorithm itself (Vellido et al., 2012). For all types of algorithms, auditing is a necessary precondition to *verify* correct functioning. For analytics algorithms with foreseeable human impact, auditing can create an ex post procedural record of complex algorithmic decision-making to unpack problematic or inaccurate decisions, or to detect discrimination or similar harms. Further work is required to design broadly applicable, low impact auditing mechanisms for algorithms (cf. Adler et al., 2016; Sandvig et al., 2014) that build upon current work in transparency and interpretability of machine learning (e.g. Kim et al., 2015; Lou et al., 2013).

All of the challenges highlighted in this review are addressable in principle. As with any technological artefact, practical solutions will require cooperation between researchers, developers and policy-makers. A final but significant area requiring further work is the translation of extant and forthcoming policy applicable to algorithms into realistic regulatory mechanisms and standards. The forthcoming EU General Data Protection Regulation (GDPR) in particular is indicative of the challenges to be faced globally in regulating algorithms.[21]

The GDPR stipulates a number of responsibilities of data controllers and rights of data subjects relevant to decision-making algorithms. Concerning the former, when undertaking profiling controllers will be required to evaluate the potential consequences of their data-processing activities via a data protection impact assessment (Art. 35(3)(a)). In addition to assessing privacy hazards, data controllers also have to communicate these risks to the persons concerned. According to Art. 13(2)(f) and 14(2)(g) data controllers are obligated to inform the data subjects about existing profiling methods, its *significance* and its *envisaged consequences*. Art. 12(1) mandates that *clear and plain language* is used to inform about these risks.[22] Further, Recital 71 states the data controllers' obligation to explain the logic of how the decision was reached. Finally, Art. 22(3) states the data controller's duty to "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests'' when automated decision-making is applied. This obligation is rather vague and opaque.

On the rights of data subjects, the GDPR generally takes a self-determination approach. Data subjects are granted a right to object to profiling methods (Art. 21) and a right not to be subject to solely automated processed individual decision-making[23] (Art. 22). In these and similar cases the person concerned either has the right to object that such methods are used or should at least have the right to ''obtain human intervention'' in order to express their views and to ''contest the decision'' (Art. 22(3)).

At first glance these provisions defer control to the data subjects and enable them to decide how their data are used. Notwithstanding that the GDPR bears great potential to improve data protection, a number of exemptions limit the rights of data subjects.[24] The GDPR can be a toothless or a powerful mechanism to protect data subjects dependent upon its eventual legal interpretation: the wording of the regulation allows either to be true. Supervisory authorities and their future judgments will determine the effectiveness of the new framework.[25] However, additional work is required in parallel to provide normative guidelines and practical mechanisms for putting the new rights and responsibilities into practice.

These are not mundane regulatory tasks. For example, the provisions highlighted above can be interpreted to mean automated decisions must be *explainable* to data subjects. Given the connectivity and dependencies of algorithms and datasets in complex information systems, and the tendency of errors and biases in data and models to be hidden over time (see 'Misguided evidence leading to bias' section), 'explainability'[26] may prove particularly disruptive for data intensive industries. Practical requirements will need to be unpacked in the future that strike an appropriate balance between data subjects' rights to be informed about the *logic* and *consequences* of profiling, and the burden imposed on data controllers. Alternatively, it may be necessary to limit automation or particular analytic methods in particular contexts to meet transparency requirements specified in the GDPR (Tutt, 2016; Zarsky, 2016). Comparable restrictions already exist in the US Credit Reporting Act, which effectively prohibits machine learning in credit scoring because reasons for the denial of credit must be made available to consumers on demand (Burrell, 2016).

Conclusion

Algorithms increasingly mediate digital life and decision-making. The work described here has made three contributions to clarify the ethical importance of this mediation: (1) a review of existing discussion of ethical aspects of algorithms; (2) a prescriptive map to organise discussion; and (3) a critical assessment of the literature to identify areas requiring further work to develop the ethics of algorithms.

The review undertaken here was primarily limited to literature explicitly discussing algorithms. As a result, much relevant work performed in related fields is only briefly touched upon, in areas such as ethics of artificial intelligence, surveillance studies, computer ethics and machine ethics.[27] While it would be ideal to summarise work in all the fields represented in the reviewed literature, and thus in any domain where algorithms are in use, the scope of such as an exercise is prohibitive. We must therefore accept that there may be gaps in coverage for topics discussed only in relation to specific types of algorithms, and not for algorithms themselves. Despite this limitation, the prescriptive map is purposefully broad and iterative to organise discussion around the ethics of algorithms, both past and future.

Discussion of a concept as complex as 'algorithm' inevitably encounters problems of abstraction or 'talking past each other' due to a failure to specify a level of abstraction (LoA) for discussion, and thus limit the relevant set of observables (Floridi, 2008). A mature 'ethics of algorithms' does not yet exist, in part because 'algorithm' as a concept describes a prohibitively broad range of software and information systems. Despite this limitation, several themes emerged from the literature that indicate how ethics can coherently be discussed when focusing on algorithms, independently of domain-specific work.

Mapping these themes onto the prescriptive framework proposed here has proven helpful to distinguish between the kinds of ethical concerns generated by algorithms, which are often muddled in the literature. Distinct epistemic and normative concerns are often treated as a cluster. This is understandable, as the different concerns are part of a web of interdependencies. Some of these interdependencies are present in the literature we reviewed, like the connection between bias and discrimination (see 'Misguided evidence leading to bias' and 'Unfair outcomes leading to discrimination' sections) or the impact of opacity on the attribution of responsibility (see 'Inscrutable evidence leading to opacity' and 'Traceability leading to moral responsibility' sections).

The proposed map brings further dependencies into focus, like the multi-faceted effect of the presence and absence of epistemic deficiencies on the ethical ramifications of algorithms. Further, the map demonstrates that solving problems at one level does not address all types of concerns; a perfectly auditable algorithmic decision, or one that is based on conclusive, scrutable and well-founded evidence, can nevertheless cause unfair and transformative effects, without obvious ways to trace blame among the network of contributing actors. Better methods to produce evidence for some actions need not rule out all forms of discrimination for example, and can even be used to discriminate more efficiently. Indeed, one may even conceive of situations where less discerning algorithms may have fewer objectionable effects.

More importantly, as already repeatedly stressed in the above overview, we cannot in principle avoid epistemic and ethical residues. Increasingly better algorithmic tools can normally be expected to rule out many obvious epistemic deficiencies, and even help us to detect well-understood ethical problems (e.g. discrimination). However, the full conceptual space of ethical challenges posed by the use of algorithms cannot be reduced to problems related to easily identified epistemic and ethical shortcomings. Aided by the map drawn here, future work should strive to make explicit the many implicit connections to algorithms in ethics and beyond.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the University of Oxford's John Fell Fund (Brent Mittelstadt), by the PETRAS IoT Hub – a EPSRC project (Brent Mittelstadt, Luciano Floridi, Mariarosaria Taddeo), and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 657017 (Patrick Allo).

Supplementary material

The supplementary files are available at http://bds.sagepub.com/content/3/2.

Notes

1. We would like to acknowledge valuable comments and feedback of the reviewers at Big Data & Society.
2. Compare with Turner (2016) on the ontology of programs.
3. For the sake of simplicity, for the remainder of the paper we will refer generically to 'algorithms' rather than constructs, implementations and configurations.
4. Tufekci seems to have a similar class of algorithms in mind in her exploration of detecting harms. She describes 'gatekeeping algorithms' as ''algorithms that do not result in simple, 'correct' answers-instead, I focus on those that are utilized as subjective decision makers'' (Tufekci, 2015: 206).
5. The distinction between supervised and unsupervised learning can be mapped onto analytics to reveal different ways humans are 'made sense of' through data. Descriptive analytics based on unsupervised learning, seeks to identify unforeseen correlations between cases to learn something about the entity or phenomenon. Here, analysis is exploratory, meaning it lacks a specific target or hypothesis. In this way, new models and classifications can be defined. In contrast, predictive analytics based on supervised learning seeks to match cases to pre-existing classes to infer knowledge about the case. Knowledge about the assigned classes is used to make predictions about the case (Van Otterlo, 2013).
6. The term 'probable knowledge' is used here in the sense of Hacking (2006) where it is associated with the emergence of probability and the rise of statistical thinking (for instance in the context of insurance) that started in the 17th Century.
7. In mainstream analytic epistemology this issue is connected to the nature of justification, and the importance of having access to one's own justifications for a specific belief (Kornblith, 2001). In the present context, however, we are concerned with a more interactive kind of justification: human agents need to be able to understand how a conclusion reached by an algorithm is justified in view of the data.
8. The often blamed opacity of algorithms can only partially explain why this is the case. Another aspect is more closely related to the role of re-use in the development of algorithms and software-artefacts; from the customary use of existing libraries, to the repurposing of existing tools and methods for different purposes (e.g. the use of seismological models of aftershocks in predictive policing (Mohler et al., 2011), and the tailoring of general tools for specific methods. Apart from the inevitable distribution of responsibilities, this highlights the complex relation between good design (the re-use philosophy promoted in *Structured Programming*) and the absence of malfunction, and reveals that even the designers of software-artefacts regularly treat part of their work as black boxes (Sametinger, 1997).
9. See Appendix 1 for information on the methodology, search terms and query results of the review.
10. A distinction must be made, however, between the ethical justifiability of acting upon mere correlation and a broader ethics of inductive reasoning which overlaps with extant critiques of statistical and quantitative methods in research. The former concerns the thresholds of evidence required to justify actions with ethical impact. The latter concerns a lack of reproducibility in analytics that distinguishes it in practice from science (cf. Feynman, 1974; Ioannidis, 2005; Vasilevsky et al., 2013) and is better understood as an issue of epistemology.
11. Introna and Nissenbaum's article (2000) is among the first publications on this topic. The article compares search engines to publishers and suggests that, like publishers, search engines filter information according to market conditions, i.e. according to consumers' tastes and preferences, and favour powerful actors. Two corrective mechanisms are suggested: embedding the ''value of fairness as well as [a] suite of values represented by the ideology of the Web as a public good'' (Introna and Nissenbaum, 2000: 182) in the design of indexing and ranking algorithms, and transparency of the algorithms used by search engines. More recently, Zarsky (2013) has

provided a framework and in-depth legal examination of transparency in predictive analytics.

12. This is a contentious claim. Bozdag (2013) suggests that human comprehension has not increased in parallel to the exponential growth of social data in recent years due to biological limitations on information processing capacities. However, this would appear to discount advances in data visualization and sorting techniques to help humans comprehend large datasets and information flows (cf. Turilli and Floridi, 2009). Biological capacities may not have increased, but the same cannot be said for tool-assisted comprehension. One's position on this turns on whether *technology-assisted* and *human* comprehension are categorically different.
13. The context of autonomous weapon systems is particularly relevant here; see Swiatek (2012).
14. The argument that technology design unavoidably value-laden is not universally accepted. Kraemer et al. (2011) provide a counterargument from the reviewed literature. For them, algorithms are value-laden only "if one cannot rationally choose between them without explicitly or implicitly taking ethical concerns into account." In other words, designers make value-judgments that express views "on how things ought to be or not to be, or what is good or bad, or desirable or undesirable" (Kraemer et al., 2011: 252). For Kraemer et al. (2011), algorithms that produce hypothetical value-judgments or recommended courses of action, such as clinical decision support systems, can be value-neutral because the judgments produced are hypothetical. This approach would suggest that autonomous algorithms are value-laden by definition, but only because the judgments produced are put into action by the algorithm. This conception of value neutrality appears to suggest that algorithms are designed in value-neutral spaces, with the designer disconnected from a social and moral context and history that inevitably influences her perceptions and decisions. It is difficult to see how this could be the case (cf. Friedman and Nissenbaum, 1996).
15. Clear sources of discrimination are not consistently identified in the reviewed literature. Barocas (2014) helpfully clarifies five possible sources of discrimination related to biased analytics: (1) inferring membership in a protected class; (2) statistical bias; (3) faulty inference; (4) overly precise inferences; and (5) shifting the sample frame.
16. Danna and Gandy (2002) provide a demonstrative example in the Royal Bank of Canada which 'nudged' customers on fee-for-service to flat-fee service packages after discovering (through mining in-house data) that customers on the latter offered greater lifetime value to the bank. Customers unwilling to move to flat-fee services faced disincentives including higher prices. Through price discrimination customers were pushed towards options reflecting the bank's interests. Customers unwilling to move were placed into a weak bargaining position in which they were 'invited to leave': losing some customers in the process of shifting the majority to more profitable flat-fee packages meant the bank lacked incentive to accommodate minority interests despite the risk of losing minority fee-for-service customers to competitors.
17. Data subjects can be considered to have a right to identity. Such a right can take many forms, but the existence of *some* right to identity is difficult to dispute. Floridi (2011) conceives of personal identity as constituted by information. Taken as such, any right to informational privacy translates to a right to identity by default, understood as the right to manage information about the self that constitutes one's identity. Hildebrandt and Koops (2010) similarly recognise a right to form identity without unreasonable external influence. Both approaches can be connected to the right to personality derived from the European Convention on Human Rights.
18. A further distinction can be made between artificial moral agents and artificial ethical agents. Artificial moral agents lack true 'artificial intelligence' or the capacity for reflection required to decide and justify an ethical course of action. Artificial ethical agents can "calculate the best action in ethical dilemmas using ethical principles" (Moor, 2006) or frameworks derived thereof. In contrast, artificial morality requires only that machines act 'as if' they are moral agents, and thus make ethically justified decisions according to pre-defined criteria (Moor, 2006). The construction of artificial morality is seen as the immediate and imminently achievable challenge for machine ethics, as it does not first require artificial intelligence (Allen et al., 2006). With that said, the question of whether "it is possible to create artificial full ethical agents" continues to occupy machine ethicists (Tonkens, 2012: 139).
19. Tonkens (2012) however argues that agents embedded with virtue-based frameworks would find their creation ethically impermissible due to the impoverished sense of virtues a machine could actually develop. In short, the character development of humans and machines are too dissimilar to compare. He predicts that unless autonomous agents are treated as full moral agents comparable to humans, existing social injustices will be exacerbated as autonomous machines are denied the freedom to express their autonomy by being forced into service of the needs of the designer. This concern points to a broader issue in machine ethics concerning whether algorithms and machines with decision-making autonomy will continue to be treated as passive tools as opposed to active (moral) agents (Wiegel and Berg, 2009).
20. Except for trivial cases, the presence of false positives and false negatives in the work of algorithms, particularly machine learning, is unavoidable.
21. It is important to note that this regulation even applies to data controllers or processors that are not established within the EU, if the monitoring (including predicting and profiling) of behaviour is focused on data subjects that are located in the EU (Art 3(2)(b) and Recital 24).
22. In cases where informed consent is required, Art. 7(2) stipulates that non-compliance with Art. 12(1) renders given consent not legally binding.
23. Recital 71 explains that solely automated individual decision-making has to be understood as a method "which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices

without any human intervention'' and includes profiling that allows to ''predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements.''

24. Art. 21(1) explains that the right to object to profiling methods can be restricted ''if the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims.'' In addition, Art. 23(1) stipulates that the rights enshrined in Art. 12 to 22 – including the right to object to automated decision-making – can be restricted in cases such as ''national security, defence; public security;(. . .); other important objectives of general public interest of the Union or of a Member State, in particular an important economic or financial interest of the Union or of a Member State, including monetary, budgetary and taxation a matters, public health and social security; (. . .); the prevention, investigation, detection and prosecution of breaches of ethics for regulated professions; (...);''. As a result, these exemptions also apply to the right to access (Art. 15 – the right to obtain information if personal data are being processed) as well as the right to be forgotten (Art. 17).
25. Art. 83(5)(b) invests supervisory authorities with the power to impose fines up to 4% of the total worldwide annual turnover in cases where rights of the data subjects (Art. 12 to 22) have been infringed. This lever can be used to enforce compliance and to enhance data protection.
26. 'Explainability' is preferred here to 'interpretability' to highlight that the explanation of a decision must be comprehensible not only to data scientists or controllers, but to the lay data subjects (or some proxy) affected by the decision.
27. The various domains of research and development described here share a common characteristic: all make use of computing algorithms. This is not, however, to suggest that complex fields such as machine ethics and surveillance studies are subsumed by the 'ethics of algorithms' label. Rather, each domain has issues which do not originate in the design and functionality of the algorithms being used. These issues would thus not be considered part of an 'ethics of algorithms', despite the inclusion of the parent field. 'Ethics of algorithms' is thus not meant to replace existing fields of enquiry, but rather to identify issues shared across a diverse number of domains stemming from the computing algorithms they use.

References

Adler P, Falk C, Friedler SA, et al. (2016) Auditing black-box models by obscuring features. *arXiv:1602.07043 [cs, stat]*. Available at: http://arxiv.org/abs/1602.07043 (accessed 5 March 2016).

Agrawal R and Srikant R (2000) Privacy-preserving data mining. *ACM Sigmod Record*. ACM, pp. 439–450. Available at: http://dl.acm.org/citation.cfm?id=335438 (accessed 20 August 2015).

Allen C, Wallach W and Smit I (2006) Why machine ethics? *Intelligent Systems, IEEE* 21(4) Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947 (accessed 1 January 2006).

Ananny M (2016) Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 93–117.

Anderson M and Anderson SL (2007) Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28(4): 15.

Anderson M and Anderson SL (2014) Toward ethical intelligent autonomous healthcare agents: A case-supported principle-based behavior paradigm. Available at: http://doc.gold.ac.uk/aisb50/AISB50-S17/AISB50-S17-Anderson-Paper.pdf (accessed 24 August 2015).

Anderson SL (2008) Asimov's 'Three Laws of Robotics' and machine metaethics. *AI and Society* 22(4): 477–493.

Applin SA and Fischer MD (2015) New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In: *2015 IEEE international symposium on technology and society (ISTAS)*. Dublin, Ireland: IEEE, pp. 1–6.

Arendt H (1971) *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press.

Barnet BA (2009) Idiomedia: The rise of personalized, aggregated content. *Continuum* 23(1): 93–99.

Barocas S (2014) Data mining and the discourse on discrimination. Available at: https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf (accessed 20 December 2015).

Barocas S and Selbst AD (2015) *Big data's disparate impact*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available at: http://papers.ssrn.com/abstract=2477899 (accessed 16 October 2015).

Bello P and Bringsjord S (2012) On how to build a moral machine. *Topoi* 32(2): 251–266.

Birrer FAJ (2005) Data mining to combat terrorism and the roots of privacy concerns. *Ethics and Information Technology* 7(4): 211–220.

Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227.

Brey P and Soraker JH (2009) *Philosophy of Computing and Information Technology*. Elsevier.

Burrell J (2016) How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

Calders T and Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2): 277–292.

Calders T, Kamiran F and Pechenizkiy M (2009) Building classifiers with independency constraints. In: *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, Miami, USA, IEEE, pp. 13–18.

Cardona B (2008) 'Healthy ageing' policies and anti-ageing ideologies and practices: On the exercise of responsibility. *Medicine, Health Care and Philosophy* 11(4): 475–483.

Coeckelbergh M (2013) E-care as craftsmanship: Virtuous work, skilled engagement, and information technology in health care. *Medicine, Health Care and Philosophy* 16(4): 807–816.

Cohen IG, Amarasingham R, Shah A, et al. (2014) The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs* 33(7): 1139–1147.

Coll S (2013) Consumption as biopower: Governing bodies with loyalty cards. *Journal of Consumer Culture* 13(3): 201–220.

Crawford K (2016) Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology & Human Values* 41(1): 77–92.

Crnkovic GD and Çürüklü B (2011) Robots: ethical by design. *Ethics and Information Technology* 14(1): 61–71.

Danna A and Gandy OH Jr (2002) All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics* 40(4): 373–386.

Datta A, Sen S and Zick Y (2016) Algorithmic transparency via quantitative input influence. In: *Proceedings of 37th IEEE symposium on security and privacy*, San Jose, USA. Available at: http://www.ieee-security.org/TC/SP2016/papers/0824a598.pdf (accessed 30 June 2016).

Davis M, Kumiega A and Van Vliet B (2013) Ethics, finance, and automation: A preliminary survey of problems in high frequency trading. *Science and Engineering Ethics* 19(3): 851–874.

de Vries K (2010) Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information Technology* 12(1): 71–85.

Diakopoulos N (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3(3): 398–415.

Diamond GA, Pollock BH and Work JW (1987) Clinician decisions and computers. *Journal of the American College of Cardiology* 9(6): 1385–1396.

Domingos P (2012) A few useful things to know about machine learning. *Communications of the ACM* 55(10): 78–87.

Dwork C, Hardt M, Pitassi T, et al. (2011) Fairness through awareness. *arXiv:1104.3913 [cs]*. Available at: http://arxiv.org/abs/1104.3913 (accessed 15 February 2016).

Elish MC (2016) Moral crumple zones: Cautionary tales in human–robot interaction (WeRobot 2016). SSRN. Available at: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id = 2757236 (accessed 30 June 2016).

European Commission (2012) *Regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Brussels: European Commission. Available at: http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (accessed 2 April 2013).

Feynman R (1974) 'Cargo cult science' – by Richard Feynman. Available at: http://neurotheory.columbia.edu/~ken/cargo_cult.html (accessed 3 September 2015).

Floridi L (2008) The method of levels of abstraction. *Minds and Machines* 18(3): 303–329.

Floridi L (2011) The informational nature of personal identity. *Minds and Machines* 21(4): 549–566.

Floridi L (2012) Big data and their epistemological challenge. *Philosophy & Technology* 25(4): 435–437.

Floridi L (2014) *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: OUP.

Floridi L and Sanders JW (2004a) On the morality of artificial agents. *Minds and Machines* 14(3). Available at: http://dl.acm.org/citation.cfm?id = 1011949.1011964 (accessed 1 August 2004).

Floridi L and Sanders JW (2004b) On the morality of artificial agents. *Minds and Machines* 14(3). Available at: http://dl.acm.org/citation.cfm?id = 1011949.1011964 (accessed 1 August 2004).

Floridi L, Fresco N and Primiero G (2014) On malfunctioning software. *Synthese* 192(4): 1199–1220.

Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14(3): 330–347.

Fule P and Roddick JF (2004) Detecting privacy and ethical sensitivity in data mining results. In: *Proceedings of the 27th Australasian conference on computer science – Volume 26*, Dunedin, New Zealand, Australian Computer Society, Inc., pp. 159–166. Available at: http://dl.acm.org/citation.cfm?id = 979942 (accessed 24 August 2015).

Gadamer HG (2004) *Truth and Method*. London: Continuum International Publishing Group.

Glenn T and Monteith S (2014) New measures of mental state and behavior based on data collected from sensors, smartphones, and the internet. *Current Psychiatry Reports* 16(12): 1–10.

Goldman E (2006) Search engine bias and the demise of search engine utopianism. *Yale Journal of Law & Technology* 8: 188–200.

Granka LA (2010) The politics of search: A decade retrospective. *The Information Society* 26(5): 364–374.

Grindrod P (2014) *Mathematical Underpinnings of Analytics: Theory and Applications*. Oxford: OUP.

Grodzinsky FS, Miller KW and Wolf MJ (2010) Developing artificial agents worthy of trust: 'Would you buy a used car from this artificial agent?' *Ethics and Information Technology* 13(1): 17–27.

Hacking I (2006) *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.

Hajian S and Domingo-Ferrer J (2013) A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering* 25(7): 1445–1459.

Hajian S, Monreale A, Pedreschi D, et al. (2012) Injecting discrimination and privacy awareness into pattern discovery. In: *Data mining workshops (ICDMW), 2012 IEEE 12th international conference on,* Brussels, Belgium, IEEE, pp. 360–369. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber = 6406463 (accessed 3 November 2015).

Hildebrandt M (2008) Defining profiling: A new type of knowledge? In: Hildebrandt M and Gutwirth S (eds) *Profiling the European Citizen* the Netherlands: Springer, pp. 17–45Available at: http://link.springer.com/chapter/10.1007/978-1-4020-6914-7_2 (accessed 14 May 2015).

Hildebrandt M (2011) Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy & Technology* 24(4): 371–390.

Hildebrandt M and Koops B-J (2010) The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review* 73(3): 428–460.

Hill RK (2015) What an algorithm is. *Philosophy & Technology* 29(1): 35–59.

Illari PM and Russo F (2014) *Causality: Philosophical Theory Meets Scientific Practice*. Oxford: Oxford University Press.

Introna LD and Nissenbaum H (2000) Shaping the Web: Why the politics of search engines matters. *The Information Society* 16(3): 169–185.

Ioannidis JPA (2005) Why most published research findings are false. *PLoS Medicine* 2(8): e124.

James G, Witten D, Hastie T, et al. (2013) *An Introduction to Statistical Learning*. Vol. 6, New York: Springer.

Johnson JA (2006) *Technology and pragmatism: From value neutrality to value criticality*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available at: http://papers.ssrn.com/abstract=2154654 (accessed 24 August 2015).

Johnson JA (2013) *Ethics of data mining and predictive analytics in higher education*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available at: http://papers.ssrn.com/abstract=2156058 (accessed 22 July 2015).

Kamiran F and Calders T (2010) Classification with no discrimination by preferential sampling. In: *Proceedings of the 19th machine learning conf. Belgium and the Netherlands*, Leuven, Belgium. Available at: http://wwwis.win.tue.nl/~tcalders/pubs/benelearn2010 (accessed 24 August 2015).

Kamishima T, Akaho S, Asoh H, et al. (2012) Considerations on fairness-aware data mining. In: *IEEE 12th International Conference on Data Mining Workshops*, Brussels, Belgium, pp. 378–385. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6406465 (accessed 3 November 2015).

Kim B, Patel K, Rostamizadeh A, et al. (2015) Scalable and interpretable data representation for high-dimensional, complex data. *AAAI*. pp. 1763–1769.

Kim H, Giacomin J and Macredie R (2014) A qualitative study of stakeholders' perspectives on the social network service environment. *International Journal of Human–Computer Interaction* 30(12): 965–976.

Kitchin R (2016) Thinking critically about and researching algorithms. *Information, Communication & Society* 20(1): 14–29.

Kornblith H (2001) *Epistemology: Internalism and Externalism*. Oxford: Blackwell.

Kraemer F, van Overveld K and Peterson M (2011) Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251–260.

Lazer D, Kennedy R, King G, et al. (2014) The parable of Google flu: Traps in big data analysis. *Science* 343(6176): 1203–1205.

Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 494–511.

Levenson JL and Pettrey L (1994) Controversial decisions regarding treatment and DNR: An algorithmic Guide for the Uncertain in Decision-Making Ethics (GUIDE). *American Journal of Critical Care: An Official Publication, American Association of Critical-Care Nurses* 3(2): 87–91.

Lewis SC and Westlund O (2015) Big data and journalism. *Digital Journalism* 3(3): 447–466.

Lomborg S and Bechmann A (2014) Using APIs for data collection on social media. *Information Society* 30(4): 256–265.

Lou Y, Caruana R, Gehrke J, et al. (2013) Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*. Chicago, USA, ACM, pp. 623–631.

Louch MO, Mainier MJ and Frketich DD (2010) An analysis of the ethics of data warehousing in the context of social networking applications and adolescents. In: *2010 ISECON Proceedings*, Vol. 27 no. 1392, Nashville, USA.

Lupton D (2014) The commodification of patient opinion: The digital patient experience economy in the age of big data. *Sociology of Health & Illness* 36(6): 856–869.

MacIntyre A (2007) *After Virtue: A Study in Moral Theory*, 3rd ed. London: Gerald Duckworth & Co Ltd. Revised edition.

Macnish K (2012) Unblinking eyes: The ethics of automating surveillance. *Ethics and Information Technology* 14(2): 151–167.

Mahajan RL, Reed J, Ramakrishnan N, et al. (2012) *Cultivating emerging and black swan technologies. ASME 2012 International Mechanical Engineering Congress and Exposition*, Houston, USA, pp. 549–557.

Markowetz A, Błaszkiewicz K, Montag C, et al. (2014) Psycho-informatics: Big data shaping modern psychometrics. *Medical Hypotheses* 82(4): 405–411.

Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that will Transform How We Live, Work and Think*. London: John Murray.

Mazoué JG (1990) Diagnosis without doctors. *Journal of Medicine and Philosophy* 15(6): 559–579.

Miller B and Record I (2013) Justified belief in a digital age: On the epistemic implications of secret Internet technologies. *Episteme* 10(2): 117–134.

Mittelstadt BD and Floridi L (2016) The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 22(2): 303–341.

Mohler GO, Short MB, Brantingham PJ, et al. (2011) Self-exciting point process modeling of crime. *Journal of the American Statistical Association* 106(493): 100–108.

Moor JH (2006) The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE* 21(4). Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667948 (accessed 1 January 2006).

Morek R (2006) Regulatory framework for online dispute resolution: A critical view. *The University of Toledo Law Review* 38: 163.

Naik G and Bhide SS (2014) Will the future of knowledge work automation transform personalized medicine? *Applied & Translational Genomics, Inaugural Issue* 3(3): 50–53.

Nakamura L (2013) *Cybertypes: Race, Ethnicity, and Identity on the Internet*. New York: Routledge.

Newell S and Marabelli M (2015) Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems* 24(1): 3–14.

Neyland D (2016) Bearing accountable witness to the ethical algorithmic system. *Science, Technology & Human Values* 41(1): 50–76.

Orseau L and Armstrong S (2016) Safely interruptible agents. Available at: http://intelligence.org/files/Interruptibility.pdf (accessed 12 September 2016).

Pariser E (2011) *The Filter Bubble: What the Internet is Hiding from You*. London: Viking.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Patterson ME and Williams DR (2002) *Collecting and Analyzing Qualitative Data: Hermeneutic Principles, Methods and Case Examples*. Advances in tourism Application Series, Champaign, IL, Champaign, USA: Sagamore Publishing, Inc. Available at: http://www.treesearch.fs.fed.us/pubs/29421 (accessed 7 November 2012).

Portmess L and Tower S (2014) Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of Big Data. *Ethics and Information Technology* 17(1): 1–9.

Raymond A (2014) The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id = 2469291 (accessed 22 July 2015).

Romei A and Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29(5): 582–638.

Rubel A and Jones KML (2014) *Student privacy in learning analytics: An information ethics perspective*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available at: http://papers.ssrn.com/abstract = 2533704 (accessed 22 July 2015).

Sametinger J (1997) *Software Engineering with Reusable Components*. Berlin: Springer Science & Business Media.

Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. Available at: http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf (accessed 13 February 2016).

Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

Shackelford SJ and Raymond AH (2014) Building the virtual courthouse: Ethical considerations for design, implementation, and regulation in the world of Odr. *Wisconsin Law Review* (3): 615–657.

Shannon CE and Weaver W (1998) *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Simon J (2010) The entanglement of trust and knowledge on the web. *Ethics and Information Technology* 12(4): 343–355.

Simon J (2015) Distributed epistemic responsibility in a hyperconnected era. In: Floridi L (ed.) *The Onlife Manifesto*. Springer International Publishing, pp. 145–159. Available at: http://link.springer.com/chapter/10.1007/978-3-319-04093-6_17 (accessed 17 June 2016).

Stark M and Fins JJ (2013) Engineering medical decisions. *Cambridge Quarterly of Healthcare Ethics* 22(4): 373–381.

Sullins JP (2006) When is a robot a moral agent? Available at: http://scholarworks.calstate.edu/xmlui/bitstream/handle/10211.1/427/Sullins%20Robots-Moral%20Agents.pdf?sequence = 1 (accessed 20 August 2015).

Sweeney L (2013) Discrimination in online ad delivery. *Queue* 11(3): 10:10–10:29.

Swiatek MS (2012) Intending to err: The ethical challenge of lethal, autonomous systems. *Ethics and Information Technology*14(4). Available at: https://www.scopus.com/inward/record.url?eid = 2-s2.0-84870680328&partnerID = 40&md5 = 018033cfd83c46292370e160d4938ffa (accessed 1 January 2012).

Taddeo M (2010) Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines* 20(2): 243–257.

Taddeo M and Floridi L (2015) The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics* 1–29.

Taylor L, Floridi L and van der Sloot B (eds) (2017) *Group Privacy: New Challenges of Data Technologies*, 1st ed. New York, NY: Springer.

Tene O and Polonetsky J (2013a) Big data for all: Privacy and user control in the age of analytics. Available at: http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle = hein.journals/nwteintp11&section = 20 (accessed 2 October 2014).

Tene O and Polonetsky J (2013b) Big Data for all: Privacy and user control in the age of analytics. Available at: http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle = hein.journals/nwteintp11&section = 20 (accessed 2 October 2014).

Tonkens R (2012) Out of character: On the creation of virtuous machines. *Ethics and Information Technology* 14(2): 137–149.

Tufekci Z (2015) Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Journal on Telecommunications and High Technology Law* 13: 203.

Turilli M (2007) Ethical protocols design. *Ethics and Information Technology* 9(1): 49–62.

Turilli M and Floridi L (2009) The ethics of information transparency. *Ethics and Information Technology* 11(2): 105–112.

Turner R (2016) The philosophy of computer science. Spring 2016. In: Zalta EN (ed.) *The Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/archives/spr2016/entries/computer-science/ (accessed 21 June 2016).

Tutt A (2016) *An FDA for algorithms*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available at: http://papers.ssrn.com/abstract=2747994 (accessed 13 April 2016).

Valiant LG (1984) A theory of the learnable. *Communications of the Journal of the ACM* 27: 1134–1142.

van den Hoven J and Rooksby E (2008) Distributive justice and the value of information: A (broadly) Rawlsian approach. In: van den Hoven J and Weckert J (eds) *Information Technology and Moral Philosophy*. Cambridge: Cambridge University Press, pp. 376–396.

Van Otterlo M (2013) A machine learning view on profiling. In: Hildebrandt M and de Vries K (eds) *Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology*. Abingdon: Routledge, pp. 41–64.

Van Wel L and Royakkers L (2004) Ethical issues in web data mining. *Ethics and Information Technology* 6(2): 129–140.

Vasilevsky NA, Brush MH, Paddock H, et al. (2013) On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ* 1: e148.

Vellido A, Martín-Guerrero JD and Lisboa PJ (2012) Making machine learning models interpretable. In*: ESANN 2012 proceedings*, Bruges, Belgium, pp. 163–172.

Wiegel V and van den Berg J (2009) Combining moral theory, modal logic and mas to create well-behaving artificial agents. *International Journal of Social Robotics* 1(3): 233–242.

Wiener N (1988) *The Human Use of Human Beings: Cybernetics and Society*. Da Capo Press.

Wiltshire TJ (2015) A prospective framework for the design of ideal artificial moral agents: Insights from the science of heroism in humans. *Minds and Machines* 25(1): 57–71.

Zarsky T (2013) Transparent predictions. *University of Illinois Law Review* 2013(4). Available at: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2324240 (accessed 17 June 2016).

Zarsky T (2016) The trouble with algorithmic decisions an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology & Human Values* 41(1): 118–132.