The Social Impact of Decision-Making Algorithms: Reviewing the Influence of Agency, Responsibility and Accountability on Trust and Blame

Dan Heaton
daniel.heaton@nottingham.ac.uk
School of Computer Science, University of Nottingham
Nottingham, United Kingdom

Jérémie Clos
jeremie.clos@nottingham.ac.uk
School of Computer Science, University of Nottingham
Nottingham, United Kingdom

Elena Nichele
elena.nichele@nottingham.ac.uk
School of Computer Science, University of Nottingham
Nottingham, United Kingdom

Joel E. Fischer
joel.fischer@nottingham.ac.uk
School of Computer Science, University of Nottingham
Nottingham, United Kingdom

ABSTRACT

Despite the ever-growing use of decision-making algorithms in daily life, there has been limited examination of how the supposed agency, responsibility and accountability of these algorithms can have impact on whether users trust them or blame them for failures. Therefore, this contribution reviews current literature relating to these concepts to synthesise present ideas around the social impact of these systems. We highlight the challenges of defining and operationalising these concepts in the context of algorithmic governance and discuss the need for more empirical research on how decision-making algorithms impact trust and blame in practice. We also foreground the importance of presumed agency and whether human agency is mitigated by increased algorithmic agency. After this, we use the AREA 4P responsible research and innovation framework to reflect on the findings of the literature review, which emphasises the need for a more nuanced understanding of the impact of agency, responsibility and accountability on trust and blame in algorithmic decision-making. By addressing these concerns and gaps in research, the authors argue that scholars can develop more effective strategies for ensuring responsible and ethical governance of decision-making algorithms.

KEYWORDS

decision-making algorithms, responsible research and innovation, trust, blame, agency, responsibility, accountability

ACM Reference Format:
Dan Heaton, Jérémie Clos, Elena Nichele, and Joel E. Fischer. 2023. The Social Impact of Decision-Making Algorithms: Reviewing the Influence of Agency, Responsibility and Accountability on Trust and Blame. In *Proceedings of July 11-12 2023 (TAS '23).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

TAS '23, Edinburgh, UK,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Decision-making algorithms are widely used in various domains, such as finance, healthcare, and criminal justice [26, 58, 63]. These algorithms aim to enhance the decision-making process, but concerns about agency can arise, usually when negative outcomes occur [14, 16]. The complexity and opacity of decision-making algorithms means it is challenging to determine who or what is responsible for negative outcomes [37, 65, 71]. Similarly, responsibility and accountability have been examined in the context of decision-making algorithms. Responsibility ensures that the system is designed and deployed ethically and with the appropriate considerations for its impact on society [37, 71]. Accountability examines who is answerable to the system's performance and outcomes, including any errors, biases, or unintended consequences [24, 72]. Some research has suggested that the responsibility and accountability for negative outcomes should be shared among all actors involved in the development and use of these algorithms, including developers, operators, data providers, regulators, and the algorithms themselves [52].

As decision-making algorithms become more ubiquitous, it is crucial to understand how they impact public perceptions of trust and blame [11, 66]. Trust is a critical component of successful implementation and adoption of decision-making algorithms, as users must have confidence in the system's ability to make unbiased and accurate decisions [2]. Conversely, the allocation of blame can occur when negative outcomes are attributed to a decision-making algorithm, which can result in a loss of trust and reluctance to use the technology [39, 75]. Several studies have explored the relationship between trust and blame in decision-making algorithms, and the factors that can influence them, including transparency [16], perceived control [37, 72] and the social context in which the system is implemented [22].

Despite the growing interest in this topic, there remains little exploration of how trust and blame in decision-making algorithms is impacted by perceived social agency, responsibility and accountability of systems. This is especially true when the algorithms impact the general public as opposed to a specific set of users, such as the NHS Covid-19 app [35]. Understanding the broader societal implications of these technologies is essential for ensuring that their development and use is socially responsible and does not exacerbate

existing inequalities [22]. Examining how agency, responsibility and accountability impact whether users of decision-making algorithms either trust them or blame them could allow developers, researchers and promoters of these systems to overcome complex barriers to adoption. By exploring the relationship between agency, responsibility, accountability, and trust and blame directed towards decision-making algorithms, this paper seeks to contribute to a better understanding of the social impact of these technologies through the process of reviewing current work relating to these concepts.

Therefore, the primary aim of this work is to review and synthesise existing literature that relates to agency, responsibility and accountability and how these concepts impact trust and blame directed towards decision-making algorithms. Alongside this, a secondary aim of this work is to use principles of Responsible Research and Innovation (RRI) to critically reflect and analyse current directions of research to offer suggestions for how upcoming research agendas into this topic could be shaped. This approach seeks to ensure that the development of decision-making algorithms is not just technically efficient, but also socially responsible, by considering the broader societal impacts of these technologies.

This review and reflection is split into four distinct sections. This first part of this contribution presents explorations of decision-making algorithms and their supposed agency, responsibility and accountability. The second part of the review then examines how this supposed agency, responsibility and accountability can feed into feelings of trust (and, thus distrust) or blame. Next, we use principles of RRI, through the AREA 4P framework, to reflect on the state of the art with regard to how agency, responsibility and accountability can impact trust and blame in decision-making algorithms. Finally, this contribution distinguishes research gaps in the current field and proposes further lines of inquiry based on the review and reflection undertaken.

2 METHODOLOGICAL APPROACH

This section details our approach to conducting our systematic review into this topic.

2.1 Search Strategy

The review followed a systematic search strategy to identify relevant studies. We searched several academic databases, including Google Scholar, using a combination of keywords related to agency, responsibility, accountability, trust and blame in decision-making algorithms. The search was limited to articles published in English between 2000 and 2023 to ensure a relevance to contemporary work was maintained.

2.2 Study Selection

The inclusion criteria for studies were:

- The study had to be published in an academic journal or conference proceedings.
- The study should focus on the impact of agency, responsibility, and accountability on trust and blame in decision-making algorithms.
- The study had to be published between 2000 and 2023 (this was exclusive of definitions of key concepts).
- Additionally, but not exclusively, the study should be related to themes concerning RRI, e.g. ethics, governance, public engagement.

The exclusion criteria were:

- Studies that did not meet the inclusion criteria.
- Studies that focused on topics other than decision-making algorithms, agency, responsibility, accountability, trust, and blame.
- Studies that were not published in English.
- Studies that were published before 2000.

Given the exclusion criteria, there is no doubt that some of the insights gained will be bias, as they come from English printed peer-reviewed work that omits grey literature.

2.3 Data Extraction and Analysis

The data extraction and analysis process followed a predefined protocol. In total, 860 publications were gathered. Literature was independently screened based on the inclusion and exclusion criteria, extracting relevant data from the studies, including the author, year of publication, research question, methodology, and key findings. The quality of the studies was assessed by two of the authors using the Critical Appraisal Skills Program checklist for systematic reviews, examining the validity, reliability, and relevance of the studies [67].

As well as reviewing the studies for the purposes already stipulated, we also analysed this literature using the AREA 4P RRI framework, adapted from Jirotka et al. [38], involving four stages: Anticipate, Reflect, Engage, and Act. The Anticipate stage involves identifying the potential ethical and social implications of decision-making algorithms. The Reflect stage involves critically analysing the implications of decision-making algorithms from different ethical perspectives. The Engage stage involves examining the potential of engaging with stakeholders to understand their perspectives on decision-making algorithms. The Act stage involves proposing actions and recommendations to address the issues identified in the review.

Overall, this review followed a rigorous methodology, including a systematic search strategy, predefined inclusion and exclusion criteria, data extraction and analysis, and quality assessment. The review applied the AREA 4P RRI framework to synthesise and analyse the existing literature on how agency, responsibility, and accountability impact trust and blame directed towards decision-making algorithms. The review findings provide valuable insights into the ethical and social implications of agency, responsibility and accountability of decision-making algorithms and inform best practices for their development and use.

3 REVIEWING AGENCY, RESPONSIBILITY AND ACCOUNTABILITY

This section of the review will focus on agency, accountability and responsibility. For the purposes of this exploration, agency refers to the ability of an entity to act on its own, without being influenced by external factors. Responsibility and accountability are related concepts, but they have different meanings. Responsibility refers to the obligation to take action or make decisions based on one's

role or position, while accountability refers to the responsibility of an individual or entity for the consequences of their actions. Each of these concepts will be explored in relation to decision-making algorithms.

3.1 Agency and Decision-Making Algorithms

Agency, for the purpose of this exploration, refers to the capacity of individuals or groups to act intentionally, make choices, and exert influence over their environment [5, 28, 77]. According to Bandura, agency involves a range of cognitive, behavioral, and motivational processes that enable individuals to set goals, develop plans, and execute actions to achieve desired outcomes [5]. Agency is also influenced by social and cultural contexts, as individuals' beliefs, values, and norms shape their understanding of the available options and the extent of their freedom to act [4, 48]. As Zimmerman argues, agency is a dynamic and interactive process that requires an ongoing negotiation between individuals and their environment, as individuals adjust their strategies and goals in response to changing circumstances [77]. Therefore, agency is not a fixed or innate characteristic of individuals, but rather a complex and dynamic process that is shaped by multiple factors, including personal attributes, social and cultural contexts, and environmental constraints.

One issue related to agency is the degree of autonomy decision-making algorithms possess. Some argue that algorithms have a degree of autonomy and agency, especially when they can learn and adapt from data inputs [14]. Bryson argues that algorithmic autonomy is not an all-or-nothing concept, but rather exists along a spectrum. At one end of the spectrum are algorithms that are highly deterministic and programmed to follow specific rules and decision-making processes, while at the other end are algorithms that are able to learn and adapt from data inputs, making decisions that are not explicitly programmed or predetermined by humans. Additionally, she highlights the potential benefits and risks associated with algorithmic autonomy. On one hand, autonomous algorithms can help improve efficiency and accuracy in a wide range of fields, from medical diagnosis to self-driving cars. On the other hand, if not properly designed and regulated, autonomous algorithms can pose significant risks, such as perpetuating bias or making decisions that harm people or society.

However, Floridi et al. argue that algorithms are not autonomous and that their outputs are the result of human design choices [26]. In their work, they contend that algorithms cannot be considered autonomous as they are created by humans and, therefore, influenced by human biases, values, and intentions. Although algorithms can be programmed to learn and adapt from data inputs, human designers are responsible for selecting the data to be used, analysing it, and deciding what to do with the outputs. The authors propose that algorithms should be used to enhance human capabilities rather than replace them, and urge us to consider the ethical implications of using algorithms in decision-making. In particular, they highlight the potential impact of algorithmic decisions on people's lives in areas like healthcare, criminal justice, and finance. Ultimately, the authors advocate for a human-centered approach to algorithm development and usage, which prioritises the common good and human values. Therefore, this suggests that agency should be attributed to human designers rather than the algorithms themselves.

The increasing use of decision-making algorithms in various domains has raised concerns about their impact on social agency, particularly in terms of how they affect the decision-making processes of individuals and the accountability of the systems and the actors involved. Some scholars have argued that decision-making algorithms may impede social agency by replacing human judgement and decision-making with automated processes. For example, Burrell highlights the problem of opacity in machine learning algorithms [16]. She argues that opaque algorithms, which are those that are difficult or impossible to interpret, can lead to blurred agency and transparency, and may perpetuate bias and discrimination. Additionally, Pasquale's analysis of the rise of algorithms and their impact on society, specifically in the realms of finance and information, is relevant to the concept of agency [58]. He argues that the opacity of these algorithms removes agency from the public, as decisions are being made without their input or understanding. This lack of transparency also limits the agency of those who are impacted by algorithmic decisions, as they have little recourse to challenge or contest those decisions.

However, others have emphasised the need to consider how algorithms can be designed to support rather than undermine social agency, and how regulation can be ensured in algorithmic decision-making. Diakopoulos examined accountability in algorithmic decision making and argues that transparency and explainability are necessary for accountability [22]. He suggests that a combination of technical and social solutions are needed to address this issue. Moreover, he stresses the importance of involving affected communities and stakeholders in the design and implementation of algorithmic decision making systems. This is similar to the findings of Selbst et al., who suggest that incorporating diverse perspectives and values into the design and implementation of sociotechnical systems can help promote fairness, which increases the agency of individuals and groups impacted by these systems [65]. Additionally, the level of detail in which data is collected and analysed impacts fairness, which can affect the agency of individuals and groups if they are not fairly represented or impacted by the system.

One of the key issues in the literature on social agency and decision-making algorithms is the role of transparency and explainability in fostering agency. Some scholars have argued that transparency is necessary to enable individuals to understand the decision-making process and to challenge algorithmic decisions if necessary [31, 76]. However, others have noted that the complexity of algorithms and the lack of transparency in their development and implementation may hinder accountability and undermine agency [16, 58].

Another important aspect of the literature on social agency and decision-making algorithms is the need for interdisciplinary approaches to address the ethical and social implications of algorithmic decision-making. Scholars from computer science, philosophy, social sciences, and humanities have emphasised the importance of considering the broader societal and political implications of algorithms, beyond their immediate technical functionality [25, 52]. Such interdisciplinary approaches can help to develop more nuanced understandings of the relationship between social agency, decision-making algorithms, and accountability, and to identify strategies for ensuring that these systems are developed and implemented in ways that support rather than undermine social agency.

To address some of the challenges, explainable artificial intelligence (XAI) has been proposed as a solution for enhancing transparency and accountability in decision-making algorithms. XAI involves designing algorithms that can provide explanations for their outputs in a human-understandable format [34]. By doing so, developers, operators, and regulators can gain better insights into how these algorithms work and why they produce certain outputs, enhancing accountability and responsibility [65].

3.2 Responsibility and Decision-Making Algorithms

Responsibility is defined as the moral and social obligations of individuals or groups to act in accordance with certain standards or norms [9, 59]. Responsibility can be understood as a combination of two key elements: attribution and accountability. Attribution refers to the recognition of one's role in a particular situation or outcome, while accountability refers to the expectation that one will take action to address or repair any harm caused [13, 59]. Responsibility is also closely linked to agency, as individuals' capacity to act intentionally and make choices is a precondition for holding them responsible for their actions [73]. However, the extent to which individuals are held responsible for their actions is also influenced by various social and cultural factors, such as norms, values, and power dynamics [3, 51]. Therefore, responsibility is not an absolute or fixed concept, but rather a dynamic and context-dependent process that is shaped by a range of individual and social factors.

When considering the responsibility that decision-making algorithms possess, there are several perspectives to consider. There have been contributions that offer frameworks for understanding, such as Tsoukias, who examined the social responsibility of algorithms in society [71]. They highlight the long-standing use of autonomous artefacts and categorises the impact of their use on data collection, manipulation, recommendation, and decision-making. The framework offered identifies challenges for decision analysts, including researchers and practitioners, and emphasises the need for a community effort to address the ethical implications of algorithmic decision-making.

Additionally, others argue that the drive towards responsible adoption of automated decision-making systems fails to take into account the complexities of human judgment and the relevance of the human ability to discern ethical cues and actions. For example, through examining the representational limitations of AI systems in discerning relevant cues and actions critical to ethical deliberations, [37] contrasts them to the twin-perspectives of pragmatism and phenomenology that provide lenses through which to unpack the human process of ethical deliberation . He concluded that a socio-technical system can only meet its moral responsibilities by attributing it directly onto the human decision maker's shoulders with full human meaningful control. This approach avoids operator hand-off and automation complacency.

There have also been studies that have paid specific attention to social responsibility, rather than responsibility in general. Social responsibility refers to the ethical and moral obligations of organisations to act in the best interests of society [18], and decision-making algorithms must also uphold these principles, particularly given the potential biases and discrimination that may result from their use [22]. As a result, there has been a growing interest in developing frameworks for ethical decision-making and the responsible use of algorithms.

One such framework is the "Fairness, Accountability, and Transparency (FAT)" framework proposed by Mittelstadt et al. [52]. This framework emphasises the importance of incorporating ethical principles into the design and implementation of algorithms, with a focus on ensuring fairness, accountability, and transparency. Specifically, the authors suggest that algorithms must be designed to avoid perpetuating or amplifying biases and discrimination, and that users must be able to understand how the algorithms work and how they arrived at their decisions.

Similarly, Selbst et al. proposed the "Sociotechnical Systems (STS)" framework, which considers the interplay between technology and social systems in promoting ethical decision-making [65]. This framework emphasises the importance of incorporating diverse perspectives and values into the design and implementation of sociotechnical systems, to promote fairness and accountability. Specifically, the authors suggest that systems must be designed to reflect the values and needs of all stakeholders, ensuring system design processes are transparent and inclusive.

Overall, the development of frameworks for ethical decision-making and the responsible use of algorithms reflects a growing recognition of the need for decision-making algorithms to be responsible to society. These frameworks highlight the importance of incorporating ethical principles and values into the design and implementation of algorithms, to promote fairness and transparency, and to ensure that these technologies are used in a responsible and socially beneficial manner.

3.3 Accountability and Decision-Making Algorithms

At its core, accountability refers to the extent to which individuals or organisations are held responsible for their actions or decisions, and the consequences that result from those actions or decisions [12]. While accountability is often associated with concepts such as transparency and control, it also has broader implications related to trust, legitimacy, and democratic governance [41, 54]. Accountability can be viewed as a mechanism for ensuring that individuals or organisations are answerable to those who are affected by their actions or decisions, and that they are held responsible for the outcomes they produce [12, 54]. This can include various forms of accountability, such as legal accountability, political accountability, and social accountability [64]. In practice, accountability is often implemented through mechanisms such as performance monitoring, evaluation, and auditing, and is seen as a key factor in promoting effective and responsible governance [41, 54].

Debate exists regarding who is accountable when algorithms do not achieve the expected outcomes. One of the USACM and EUACM devised principles for algorithmic fairness is accountability, which ensures those who deploy an algorithm cannot eschew responsibility for its actions, therefore not deflecting responsibility to an automated system [27]. Despite this, research suggests many individuals and groups do shift responsibility from humans if an algorithm is involved with a decision-making process. For example, Turton et al. stated that Google and Meta deflect responsibility onto

their social media algorithms despite being in control of their own code [72].

Feier et al. looked at whether an agent is systematically judged differently when the agent is artificial rather than human [24]. They found decision-makers can actually rid themselves of guilt more easily by delegating to machines than by delegating to other people, thus showing the availability of artificial agents could provide stronger incentives for decision makers to delegate morally sensitive decisions. Therefore, it could be interpreted that decision-making algorithms are used to deflect accountability from human decision-makers.

Similarly, Bucher et al. coined the term 'algorithmic imaginary' - the way in which people imagine, perceive and experience algorithms and what these imaginations make possible [15]. This has been applied in many contexts - most suitably for this strand of research by Benjamin et al., who recently studied the response to the Ofqual A Level algorithm on Twitter through the examination of the "fuck the algorithm" chant as an imaginary of resistance to confront power in sociotechnical systems [10] . Their analysis argued that this chant made algorithms more visible to the public and prompted questions about social algorithms that shape the lives of people everyday.

The study by Burrell, discussed earlier, is relevant to accountability also [16]. To address issues with agency, she suggests designers and developers of machine learning algorithms need to take steps to increase transparency, including developing tools for auditing algorithms and making their workings more transparent to users.

Overall, these studies demonstrate the existing work on how humans attribute responsibility to decision-making algorithms socially and could pave the way for further investigation into how these algorithms could influence everyday life when accountability is placed solely on them. The seeming removal of autonomy and accountability from the human or humans that devise these algorithms could be replicated in online discourses or perceived in another way. In order for creators and promoters of these systems to be successful, having accurate insights into the current perceptions of these algorithms is important so potential misleading information can be combated.

4 INFLUENCING TRUST AND BLAME

The concepts of trust and blame in the context of decision-making algorithms are becoming staple topics of autonomous system literature. Trust is a crucial factor in ensuring ethical and responsible use of algorithms, as it is essential for users to trust that algorithms produce accurate and reliable outcomes [49]. Developers, operators, or other actors involved in the development and use of these algorithms are often held accountable when blame is assigned [26]. While trust and blame may seem contradictory, they can coexist [8] - for example, when users have confidence in the overall integrity of the algorithms while still holding developers and operators accountable for negative outcomes. The following sub-sections explore how agency, responsibility and accountability can impact the trust in – and the blaming of – decision-making algorithms in existing literature.

4.1 Trust and Decision-Making Algorithms

Trust, as a multi-dimensional construct, has been extensively studied in multiple fields, including sociology, psychology, economics, and management. Scholars have examined trust in diverse contexts, such as interpersonal relationships, organisational settings, and cross-cultural interactions [23, 49]. Trust is not only a cognitive process, but also a dynamic process influenced by individual differences, context, and various factors such as conflict, power dynamics, and external events [50]. The concept of trust has many different definitions and interpretations and there is currently no uniformed or universally agreed definition [1]. For this review, the epistemological stance undertaken will be that trust is a socially constructed concept created within an individual internally [74] as a result of interaction and experience [33]. The process of building and maintaining trust involves communication, mutual exchange, and negotiation [21]. Furthermore, trust is shaped by an individual's experiences, cultural background, and context, and is considered a socially constructed concept [29, 44].

The successful adoption and deployment of decision-making algorithms depend on the level of trust users have in them, which is influenced by the concepts of agency, responsibility, and blame. Studies have shown that users are more likely to trust algorithms that operate autonomously and produce reliable outcomes [11, 26]. In contrast, if algorithms are perceived as being influenced by external factors, such as human biases, their trustworthiness may be questioned [11].

The fulfillment of responsibilities and accountability of actors involved in the development and use of decision-making algorithms also affects trust. When developers and operators fulfill their obligations and are held accountable for their actions, users may have greater trust in the overall integrity and reliability of the algorithms. In contrast, failure to fulfill these obligations and responsibilities may lead to blame being assigned and may reduce users' trust in the algorithms [11].

More specifically when working with decision-making algorithms, Shahrdad et al. examined trust because of the exponential growth of the use of these systems in daily life, with the intention to review existing literature [66]. Prior studies indicated trust towards fully autonomous and semi-autonomous systems — such as home service robots and flight management systems — is low [46, 53]. As a result of examining these studies, alongside others, they found that managing trust will affect the development of future acceptance and adoption of these systems.

Moreover, Lyons et al. studied the verification and validation of similar decision-making algorithms and created a novel method to certify trust in them [45]. They argue that 'transparency facets' — an established communication channel between the designer, tester and user -– will enable the user to understand the goals of the system to verify its trustworthiness. Similarly, Kwiatkowska and Lahijanian called for the channels of communication to be re-examined to improve the perception of trustworthiness of these decision-making algorithms [42]. A necessity to advance the role of social trust within Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) underpinned this theory. This accounts for competence, disposition, dependence and fulfilment. Ultimately,

although both studies were inconclusive and called for more investigation, they highlight the importance of user feedback in the design and evaluation stages of decision-making algorithms creation and curation.

Additionally, Alaieri and Vellino argue that in order for people to trust these machines, their ethical principles must be transparent and predictable [2]. However, the autonomy and self-learning abilities of some robots may make their decisions non-predictable and difficult to explain, leading to increased responsibility but also decreased trust. Thus, they reinforce the need for further research and development in this area to ensure the ethical justification and trustworthiness of autonomous systems.

Despite the extensive research on trust in various contexts, there are still gaps and limitations in the literature related to how agency, responsibility, and accountability impact trust in decision-making algorithms. Firstly, there is a lack of consensus on the definition of trust, which may lead to different interpretations and inconsistent findings. Moreover, trust is a complex and dynamic concept influenced by many individual and contextual factors that may not be fully understood or controlled. Additionally, the studies reviewed in this article provide some insight into how trust is impacted by agency, responsibility, and accountability, but more research is needed to fully understand the mechanisms involved and how to design and evaluate decision-making algorithms that foster trust. Specifically, there is a need for further investigation into the role of communication, user feedback, and ethical principles in building and maintaining trust in these systems. Finally, the ethical implications of decision-making algorithms, especially in relation to autonomy and self-learning abilities, require further attention to ensure their trustworthiness.

4.2 Blame and Decision-Making Algorithms

Blame can be defined as the assignment of responsibility for a particular event or outcome, often with a negative connotation [19]. Blame can be directed towards individuals or groups, and can have various functions, such as expressing disapproval, holding individuals accountable, or seeking to assign causality [7, 19]. Blame is often accompanied by moral judgments, as it involves the evaluation of individuals' actions or omissions against certain norms or standards [62, 70]. However, the process of blaming is also influenced by various cognitive and motivational biases, such as the fundamental attribution error, which involves overestimating the role of dispositional factors and underestimating situational factors in explaining behavior [30, 61]. Therefore, blame is a complex and multifaceted process that involves a range of cognitive, emotional, and social factors, and can have significant implications for individuals' self-esteem, social relationships, and sense of justice.

The relationship between agency, responsibility, and accountability in decision-making algorithms and their impact on blame assignment has been explored. Some scholars have noted that high levels of agency in algorithms can lead to a reduction in accountability and make it difficult to assign blame for negative outcomes. For example, Mittelstadt et al. found that algorithms with a high degree of agency can result in a "responsibility gap," where neither the developers nor the algorithms are fully responsible for the outcomes produced [52]. This study also emphasised the need to determine who should be held responsible for negative outcomes in decision-making algorithms [52]. They found that, while developers and operators are typically seen as the most obvious targets of blame, others argue that blame can be shared among all actors involved in the development and use of these algorithms.

Similarly, the fulfillment of ethical responsibilities by developers and operators is also an important aspect of blame. Jobin et al. found that fulfilling ethical responsibilities can increase user trust in algorithms, while failure to do so can lead to decreased trust and increased blame assignment [39]. Similarly, Whittlestone et al. argue that ensuring ethical use of algorithms by fulfilling responsibilities is crucial for maintaining trust in technology and avoiding negative societal impacts [75].

Additionally, accountability is another key factor in the assignment of blame in decision-making algorithms. The accountability of developers and operators for the outcomes produced by algorithms they develop and use is necessary to ensure ethical and responsible use of technology. Taddeo and Floridi argue that accountability is essential for holding developers and operators responsible for the ethical use of algorithms and building user trust in technology [69]. However, blame is complex and depends on factors such as the degree of intention behind the actions, as noted by Coeckelbergh [20].

The increasing use of decision-making algorithms has led scholars to grapple with the question of how to assign blame in cases where these algorithms produce negative outcomes. Jobin et al. found that algorithms themselves can also be viewed as objects of blame, given that they may perpetuate biases or produce negative outcomes due to the design of the system [39]. However, they note that assigning blame can be challenging due to the complexity and opacity of these systems. On the other hand, Burrell argues that assigning blame is still important to ensure that decision-making algorithms are used ethically and responsibly [16]. Nonetheless, the assignment of blame is complicated by the involvement of multiple actors, including developers, operators, data providers, regulators, and the algorithms themselves [52].

Some research has been done to try and address these aforementioned challenges, Selbst et al. proposed incorporating fairness and abstraction in sociotechnical systems to ensure ethical use of decision-making algorithms [65], while Barocas and Selbst discussed the concept of disparate impact in big data [6]. Additionally, Gunning (2019) emphasised the importance of explainable artificial intelligence (XAI) to ensure transparency and accountability in decision-making algorithms [34].

However, there are still several research gaps regarding the blame assignment process. Firstly, current research lacks discussion on the cultural and societal factors that impact blame assignment. It acknowledges that cultural differences can significantly influence how blame is assigned and that emotions such as anger or fear can lead to biased decision-making. Secondly, as this review has focused solely on blame assignment within the context of decision-making algorithms, it has not fully explored the roles of other actors, such as regulators and data providers, in the process. Examining their contributions and responsibilities can provide greater insight into how blame is assigned. Thirdly, the literature reviewed does not delve into the influence of legal and ethical frameworks on blame

assignment, which could shape the ethical landscape of the field and affect how blame is assigned.

In summary, the challenge of assigning blame in the context of decision-making algorithms is multifaceted, involving not only the developers and operators but also data providers, regulators, and the algorithms themselves. Incorporating transparency, accountability, fairness, and explainability in decision-making algorithms can promote ethical and responsible use and provide clearer guidelines for assigning responsibility in cases where negative outcomes occur.

5 RESPONSIBLE RESEARCH AND INNOVATION REFLECTIONS

5.1 Background to Responsible Research and Innovation

The AREA 4P RRI framework is a model that guides ethical and responsible research and innovation (RRI) practices. It encompasses four dimensions: Anticipation, Reflection, Engagement, and Action (AREA). Anticipation involves identifying potential social impacts and ethical implications of research and innovation, while Reflection involves critically reflecting on the ethical, social, and political aspects of research and innovation. Engagement involves involving relevant stakeholders in the research and innovation process, while Action involves taking concrete actions to address any potential negative social impacts or ethical issues. The framework has been proposed as a useful tool for promoting responsible innovation in various fields, including biotechnology and nanotechnology [57, 68].

The AREA 4P RRI framework has been applied in various fields to guide ethical and responsible research practices. For instance, Ramirez-Andreotta et al. applied the framework to guide a study on the potential environmental and social impacts of a new technology for detecting heavy metals in soil [60]. Similarly, the framework has been used to guide health research, including cancer research [43], finding that using RRI-based strategies promoted collaborative processes that leveraged innovation for sustainable health systems. These applications demonstrate that the AREA 4P RRI framework is a useful tool for promoting ethical and responsible research and innovation practices. It helps researchers anticipate potential social impacts, reflect on ethical and political implications, engage with relevant stakeholders, and take actions to address any negative impacts. By doing so, the framework can help ensure that research and innovation are conducted in a socially beneficial manner.

5.2 Reflecting Using RRI

5.2.1 Anticipate. As discussed in the previous sections, decision-making algorithms are increasingly being used in various domains, including healthcare, criminal justice, finance, and employment, among others. While these algorithms have the potential to improve decision-making processes and outcomes, there are concerns about their impact on trust and blame directed towards their decisions. In this section, we anticipate the impacts of agency, responsibility, and accountability on trust and blame directed towards decision-making algorithms.

The concept of agency refers to the capacity of an entity to act and make decisions that affect others. Decision-making algorithms, by their nature, have agency as they are programmed to make decisions that affect individuals and society. However, the question arises as to who or what should be held accountable for their decisions. If the algorithm makes an erroneous decision, should the algorithm or the programmer be held accountable? In the literature reviewed, it has been argued that the responsibility for decision-making algorithms lies with both the developers and the users of the algorithm [52].

The issue of responsibility is closely related to accountability, which refers to the obligation of an entity to explain and justify its actions. The literature reviewed suggests that decision-making algorithms are often viewed as a "black box," making it difficult to determine how decisions were made and who is accountable for them [65]. This lack of transparency can erode trust in decision-making algorithms and lead to blame being directed towards the wrong parties.

The impact of agency, responsibility, and accountability on trust and blame directed towards decision-making algorithms can have significant economic, social, and environmental implications. For example, in the domain of employment, algorithms used to screen job applicants may inadvertently perpetuate biases and discrimination [6]. This can have social and economic impacts, as it can lead to qualified candidates being overlooked and exacerbate existing inequalities. Similarly, in healthcare, algorithms used to predict patient outcomes may inadvertently create disparities in healthcare access and quality, resulting in sub-optimal health outcomes [63].

To summarise, by examining works relating to how agency, responsibility, and accountability can impact trust and blame directed towards decision-making algorithms, it is clear that issues are complex and multifaceted. While decision-making algorithms have the potential to improve processes and outcomes, they also raise concerns about transparency, accountability and responsibility. To ensure that decision-making algorithms are used ethically and effectively, it is crucial to consider their impact on trust and blame and the potential economic, social and environmental implications that may arise.

5.2.2 Reflect. The previous section has shown how agency, responsibility, and accountability impact the trust and blame directed towards decision-making algorithms. The research has revealed the complexity of the issue and the need for a deeper understanding of the underlying factors that affect the use of such algorithms in decision-making processes. As a result, it is important to reflect on the purposes of, motivations for, and potential implications of examining this area of research, and the associated uncertainties, assumptions, questions, dilemmas, and social transformations these may bring.

One of the main purposes of examining this research is to improve the accountability and transparency of decision-making algorithms. By understanding the factors that affect the trust and blame directed towards these algorithms, it is possible to design systems that are more openly trustworthy and accountable, removing the 'black box'. This can help to build public trust in decision-making processes and ensure that decisions are made fairly and transparently. Additionally, examining this area of research can help to

identify potential biases and other issues that may arise from the use of decision-making algorithms, and develop strategies to address these issues.

However, examining this area of research also raises important questions and dilemmas. For example, it is unclear how much responsibility decision-making algorithms should have in the decisions they make. Should algorithms be treated as moral agents that can be held responsible for their actions, or should responsibility lie solely with the humans who design and use them? This requires further investigation and discussion. Additionally, examining this area of research may also raise questions about the broader societal implications of relying on algorithms to make decisions. For example, what are the economic, social, and environmental impacts of using decision-making algorithms, and how can these be mitigated?

Furthermore, there are uncertainties associated with this area of research. For example, there is still much we do not know about the factors that influence the trust and blame directed towards decision-making algorithms, and how these factors vary across different contexts and cultures. For instance, in collectivist cultures where trust in authority and societal harmony are highly valued, the perception of decision-making algorithms may be influenced by the perceived trustworthiness of the governing institutions and the degree of transparency in algorithmic decision-making processes [47]. On the other hand, in individualistic cultures that emphasize personal autonomy and individual rights, the focus may be more on the fairness, accountability, and explainability of the algorithms [36]. Additionally, there is a need for more research on the ethical and legal implications of using decision-making algorithms, and how these implications vary across different domains and applications.

Overall, examining the impact of agency, responsibility, and accountability on trust and blame directed towards decision-making algorithms is a complex and multifaceted issue that requires further investigation and discussion. By reflecting on the purposes, motivations, and potential implications of this area of research, this will enable us, and other researchers, to develop a better understanding of the underlying factors that affect the use of decision-making algorithms, and formulate strategies to ensure that these algorithms are trustworthy, transparent, and accountable.

5.2.3 Engage. As the conversation on the impacts of agency, responsibility, and accountability on trust and blame towards decision-making algorithms continues, it is clear that a more inclusive approach is necessary to fully understand and address these issues. Engaging with a range of stakeholders is crucial to ensure that the impacts of these systems are understood and addressed in an equitable way.

Policymakers are one key stakeholder in this process. They play a critical role in regulating the development and deployment of decision-making algorithms by establishing legal frameworks that ensure accountability and transparency. These frameworks can help increase trust in these systems, which is essential for their successful adoption [32]. Policymakers can also facilitate public dialogue and participation, which is important for creating opportunities for diverse voices to be heard in the decision-making process [56].

Developers of decision-making algorithms also have a significant role to play. They can design algorithms that are more transparent, explainable, and accountable, which can help increase trust and mitigate the potential negative consequences of these systems [52]. By taking into account the needs and perspectives of all stakeholders, including end-users and affected communities, developers can design algorithms that are fair, equitable, and beneficial to all.

End-users of decision-making algorithms also need to be actively engaged in the discussion. They are the individuals who interact with these systems and are directly impacted by their decisions. Their perspectives are critical in shaping the design and deployment of these systems, ensuring that they are transparent, accountable, and equitable [52].

Finally, affected communities, particularly those who are disproportionately impacted by these systems, must also be involved in the conversation. These communities often have unique insights into the potential impacts of decision-making algorithms and can provide valuable feedback on how to address them [65].

In summary, broader engagement and deliberation are essential to ensure that the impacts of agency, responsibility, and accountability on trust and blame towards decision-making algorithms are understood and addressed in an inclusive way. Policymakers, developers, end-users, and affected communities all have a crucial role to play in this process. By engaging these stakeholders and incorporating their perspectives, we can create decision-making algorithms that are more transparent, accountable, and equitable.

5.2.4 Act. This stage involves proposing actions and recommendations to address the issues identified in the literature review. Based on the reflection and review, the following actions and recommendations can be proposed:

(1) Enhance transparency and accountability: To promote trust in decision-making algorithms, and, thus, avoid blame, developers and operators can provide transparency and accountability in their use. This can be achieved by providing clear explanations of how algorithms work and the data used to train them, as well as taking responsibility for negative outcomes. In turn, this may shift the perception of agency from the algorithm itself to the developer or operator, ensuring a more trustworthy process.
(2) Encourage ethical use of decision-making algorithms: The ethical use of decision-making algorithms can be encouraged by setting clear ethical guidelines and standards for their development and use. Developers and operators could be trained on ethical principles and best practices, and regular audits should be conducted to ensure that algorithms are being used in an ethical and responsible manner.
(3) Foster collaboration: Collaboration among developers, operators, data providers, regulators, and other stakeholders can facilitate the development of ethical decision-making algorithms. Collaboration can lead to the sharing of knowledge, resources, and best practices, and can help ensure that ethical considerations are embedded in the development and use of algorithms.
(4) Incorporate XAI: Explainable AI (XAI) can be incorporated into decision-making algorithms to increase transparency and accountability. XAI techniques can help users understand how algorithms arrive at decisions and identify any biases or errors in the process.

(5) Address legal and ethical frameworks: Legal and ethical frameworks surrounding the development and use of decision-making algorithms can be reviewed and updated regularly to ensure they reflect changing societal values and concerns. This may help ensure that decision-making algorithms are used ethically and responsibly. As a result, blame might assigned appropriately when negative outcomes occur.
(6) Account for cultural and societal factors: Cultural and societal factors should be taken into account when assigning blame in decision-making algorithms. This can be achieved by considering the perspectives of diverse stakeholders and conducting research on how cultural and societal factors can influence the blame assignment process.
(7) Further research: Further research is needed to address gaps in the literature, such as the roles of regulators and data providers in the blame assignment process and the influence of legal and ethical frameworks on trust and blame assignment. Additionally, more research investigating how emotions and cultural factors may help influence this process.

6 OVERARCHING RESEARCH GAPS AND FUTURE WORK

This section details the overarching research gaps and ideas for future work, combining ideas from the literature reviewed and our own RRI-based reflections.

Significant attention has been paid to the impact of decision-making algorithms on trust and blame, but there are still several gaps in the literature concerning the role of agency, responsibility, and accountability in shaping these dynamics [6, 52]. One key gap is the lack of consensus on how to define and operationalise these concepts in the context of algorithmic decision-making. For example, while accountability is often discussed as a key principle in ensuring responsible algorithmic governance, there is still significant debate over who should be held accountable for algorithmic decisions and how accountability should be enforced [22, 27].

Another gap in the literature is the need for more empirical research on how decision-making algorithms impact trust and blame in practice. While there have been numerous studies examining public attitudes towards algorithmic decision-making, few have explored how these attitudes translate into actual behavior or decision-making processes [40]. Additionally, there is a need for more research on how different factors, such as the design of algorithms, the specific context in which they are used, and the characteristics of the individuals involved, impact the relationship between agency, responsibility, accountability, trust, and blame in algorithmic decision-making [17, 52].

Finally, there is a need for more interdisciplinary research that brings together perspectives from computer science, social sciences, and humanities to better understand the complex ethical, social, and political implications of algorithmic decision-making [25, 55]. This includes exploring the different ways in which algorithms can shape power dynamics, the role of transparency and explainability in building trust and accountability, and the broader societal implications of algorithmic decision-making beyond specific applications or use cases [65, 76]. By addressing these research gaps, scholars can gain a more nuanced understanding of the impact of agency, responsibility, and accountability on trust and blame in algorithmic decision-making and develop more effective strategies for ensuring responsible and ethical governance of these systems.

7 CONCLUSION

In conclusion, the literature review and RRI-based reflection provide a comprehensive overview of the impact of decision-making algorithms on trust and blame, highlighting the importance of accountability, transparency, and ethical use of algorithms. The review identified several gaps in the literature concerning the role of agency, responsibility, and accountability in shaping these dynamics. There is a lack of consensus on how to define and make use of these concepts in the context of algorithmic decision-making, as well as a need for more empirical research on how algorithms impact trust and blame in practice. Additionally, there is a need for more interdisciplinary research that brings together perspectives from computer science, social sciences, and humanities to better understand the complex ethical, social, and political implications of algorithmic decision-making.

To address these gaps, scholars can propose several actions and recommendations, as identified through the AREA 4P RRI analysis. These include enhancing transparency and accountability, encouraging ethical use of algorithms, fostering collaboration among stakeholders, incorporating XAI, updating legal and ethical frameworks, accounting for cultural and societal factors, and conducting further research. These recommendations highlight the importance of building a responsible and ethical governance structure for decision-making algorithms, one that prioritises the needs and perspectives of diverse stakeholders, and ensures that algorithms are used in an ethical and responsible manner.

Furthermore, the review emphasises that there is a need for a more nuanced understanding of the impact of agency, responsibility, and accountability on trust and blame in algorithmic decision-making. Factors such as the design of algorithms, the specific context in which they are used, and the characteristics of the individuals involved can all impact the relationship between these concepts. Therefore, scholars need to explore these factors through empirical research to develop more effective strategies for ensuring responsible and ethical governance of these systems. The interdisciplinary approach to research advocated in this review will help identify and address the broader societal implications of algorithmic decision-making beyond specific applications or use cases.

Overall, the literature review and RRI-based reflection highlight the importance of building a responsible and ethical governance structure for decision-making algorithms. This can be achieved through enhancing transparency and accountability, encouraging ethical use of algorithms, fostering collaboration among stakeholders, incorporating XAI, updating legal and ethical frameworks, accounting for cultural and societal factors, and conducting further research. By addressing these gaps, with a focus on agency, responsibility and accountability, researchers can ensure that decision-making algorithms are used in an ethical and responsible manner, and that trust in these systems is improved.

ACKNOWLEDGMENTS

All authors are supported by the UKRI Trustworthy Autonomous Systems Hub (UKRI Grant No. EP/V00784X/1). Dan Heaton is supported by the Horizon Centre for Doctoral Training at the University of Nottingham (UKRI Grant No. EP/S023305/1).

REFERENCES

[1] Barbara D. Adams, Lora E. Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol McCann. 2003. Trust in automated systems. *Toronto: Ministry of National Defence* (2003).

[2] Fahad Alaieri and André Vellino. 2016. Ethical Decision Making in Robots: Autonomy, Trust and Responsibility: Autonomy Trust and Responsibility. In *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*. Springer, 159–168.

[3] Margaret S Archer and Margaret Scotford Archer. 2000. *Being human: The problem of agency*. Cambridge University Press.

[4] Jeffrey J Arnett. 2016. The neglected 95%: why American psychology needs to become less American. (2016).

[5] Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology* 52, 1 (2001), 1–26.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review* (2016), 671–732.

[7] Roy F Baumeister. 1996. *Evil: Inside human cruelty and violence*. WH Freeman/Times Books/Henry Holt & Co.

[8] Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology* 5, 4 (2001), 323–370.

[9] Roy F Baumeister and Mark R Leary. 2017. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Interpersonal development* (2017), 57–89.

[10] Garfield Benjamin. 2022. # FuckTheAlgorithm: algorithmic imaginaries and political resistance. In *ACM Conference on Fairness, Accountability, and Transparency 2022*.

[11] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[12] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework 1. *European law journal* 13, 4 (2007), 447–468.

[13] Marcus Alphons Petrus Bovens, Mark Bovens, et al. 1998. *The quest for responsibility: Accountability and citizenship in complex organisations*. Cambridge university press.

[14] Joanna J Bryson. 2020. The artificial intelligence of the ethics of artificial intelligence. *The Oxford handbook of ethics of AI* (2020), 1.

[15] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, communication & society* 20, 1 (2017), 30–44.

[16] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016), 2053951715622512.

[17] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239.

[18] Archie B Carroll. 1979. A three-dimensional conceptual model of corporate performance. *Academy of management review* 4, 4 (1979), 497–505.

[19] D Justin Coates and Neal A Tognazzini. 2013. *Blame: Its nature and norms*. Oxford University Press on Demand.

[20] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics* 26, 4 (2020), 2051–2068.

[21] Karen S Cook and Karen A Hegtvedt. 1983. Distributive justice, equity, and equality. *Annual review of sociology* 9, 1 (1983), 217–241.

[22] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[23] Kurt T Dirks and Donald L Ferrin. 2002. Trust in leadership: meta-analytic findings and implications for research and practice. *Journal of applied psychology* 87, 4 (2002), 611.

[24] Till Feier, Jan Gogoll, and Matthias Uhl. 2021. Hiding Behind Machines: When Blame Is Shifted to Artificial Agents. *CoRR* abs/2101.11465 (2021). arXiv:2101.11465 https://arxiv.org/abs/2101.11465

[25] Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design* (2022), 535–545.

[26] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines* 28 (2018), 689–707.

[27] Simson Garfinkel, Jeanna Matthews, Stuart S Shapiro, and Jonathan M Smith. 2017. Toward algorithmic transparency and accountability. , 5–5 pages.

[28] Anthony Giddens. 1986. *The constitution of society: Outline of the theory of structuration*. Vol. 349. Univ of California Press.

[29] Anthony Giddens. 2007. The consequences of modernity. 1990. (2007).

[30] Daniel T Gilbert and Patrick S Malone. 1995. The correspondence bias. *Psychological bulletin* 117, 1 (1995), 21.

[31] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167, 2014 (2014), 167.

[32] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[33] Melanie C. Green. 2007. Trust and social interaction on the Internet. *The Oxford handbook of Internet psychology* 56 (2007), 43–51.

[34] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science robotics* 4, 37 (2019), eaay7120.

[35] Dan Heaton, Jeremie Clos, Elena Nichele, and Joel Fischer. 2023. Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. *PeerJ Computer Science* 9 (2023), e1211.

[36] Lambert Hogenhout. 2021. A framework for ethical AI at the United Nations. *arXiv preprint arXiv:2104.12547* (2021).

[37] W David Holford. 2022. 'Design-for-responsible'algorithmic decision-making systems: a question of ethical judgement and human meaningful control. *AI and Ethics* 2, 4 (2022), 827–836.

[38] Marina Jirotka, Barbara Grimpe, Bernd Stahl, Grace Eden, and Mark Hartswood. 2017. Responsible research and innovation in the digital age. *Commun. ACM* 60, 5 (2017), 62–68.

[39] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[40] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13, 3 (2020), 795–848.

[41] Jonathan GS Koppell. 2005. Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder". *Public administration review* 65, 1 (2005), 94–108.

[42] Marta Kwiatkowska and Morteza Lahijanian. 2016. Social trust: a major challenge for the future of autonomous systems. (2016).

[43] Pascale Lehoux, Hudson P Silva, Robson Rocha de Oliveira, Renata P Sabio, and Kathy Malas. 2022. Responsible innovation in health and health system sustainability: Insights from health innovators' views and practices. *Health Services Management Research* 35, 4 (2022), 196–205.

[44] N Luhmann. 1979. Trust and Power (John A. Wiley and Sons, Chichester).

[45] Joseph B. Lyons, Matthew A. Clark, Alan R. Wagner, and Matthew J. Schuelke. 2017. Certifiable trust in autonomous systems: Making the intractable tangible. *AI Magazine* 38, 3 (2017), 37–49.

[46] Poornima Madhavan and Douglas A. Wiegmann. 2007. Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 277–301.

[47] Peter Mantello, Manh-Tung Ho, Minh-Hoang Nguyen, and Quan-Hoang Vuong. 2023. Bosses without a heart: socio-demographic and cross-cultural determinants of attitude toward Emotional AI in the workplace. *AI & society* 38, 1 (2023), 97–119.

[48] Ivana Marková. 2003. *Dialogicality and social representations: The dynamics of mind*. Cambridge University Press.

[49] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[50] Debra Meyerson, Karl E Weick, Roderick M Kramer, et al. 1996. Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and research* 166 (1996), 195.

[51] Dale T Miller. 2001. Disrespect and the experience of injustice. *Annual review of psychology* 52, 1 (2001), 527–553.

[52] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.

[53] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies* 27, 5-6 (1987), 527–539.

[54] Richard Mulgan. 2000. 'Accountability': an ever-expanding concept? *Public administration* 78, 3 (2000), 555–573.

[55] Helen Nissenbaum. 2009. Privacy in context. In *Privacy in Context*. Stanford University Press.

[56] Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[57] Richard Owen, Phil Macnaghten, and Jack Stilgoe. 2020. Responsible research and innovation: From science in society to science for society, with society. In *Emerging technologies: ethics, law and governance*. Routledge, 117–126.

[58] Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information.* Harvard University Press.
[59] Philip Pettit. 2001. *A theory of freedom: from the psychology to the politics of agency.* Oxford University Press on Demand.
[60] Monica D Ramirez-Andreotta, Julia Green Brody, Nathan Lothrop, Miranda Loh, Paloma I Beamer, and Phil Brown. 2016. Improving environmental health literacy and justice through environmental exposure results communication. *International journal of environmental research and public health* 13, 7 (2016), 690.
[61] Lee Ross. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology.* Vol. 10. Elsevier, 173–220.
[62] Paul Rozin, Maureen Markwith, and Caryn Stoess. 1997. Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological science* 8, 2 (1997), 67–73.
[63] Gabrielle Samuel, SL Roberts, A Fiske, F Lucivero, S McLennan, A Phillips, S Hayes, and SB Johnson. 2022. COVID-19 contact tracing apps: UK public perceptions. *Critical Public Health* 32, 1 (2022), 31–43.
[64] Andreas Schedler, Larry Jay Diamond, and Marc F Plattner. 1999. *The self-restraining state: power and accountability in new democracies.* Lynne Rienner Publishers.
[65] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency.* 59–68.
[66] Mehdi Shahrdad and Mehdi Chehel Amirani. 2018. Detection of preterm labor by partitioning and clustering the EHG signal. *Biomedical Signal Processing and Control* 45 (2018), 109–116.
[67] Jatinder Singh. 2013. Critical appraisal skills programme. *Journal of pharmacology and Pharmacotherapeutics* 4, 1 (2013), 76–76.
[68] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2020. Developing a framework for responsible innovation. In *The ethics of nanotechnology, geoengineering and clean energy.* Routledge, 347–359.
[69] Mariarosaria Taddeo and Luciano Floridi. 2018. How AI can be a force for good. *Science* 361, 6404 (2018), 751–752.
[70] Philip E Tetlock. 1992. The impact of accountability on judgment and choice: Toward a social contingency model. In *Advances in experimental social psychology.* Vol. 25. Elsevier, 331–376.
[71] Alexis Tsoukias. 2021. Social responsibility of algorithms: an overview. *EURO Working Group on DSS: A Tour of the DSS Developments Over the Last 30 Years* (2021), 153–166.
[72] William Turton. 2017. The algorithm is innocent. Google and Facebook deflect responsibility onto algorithms, as if they don't control their own code. *The Outline* (2017).
[73] R Jay Wallace. 1998. Responsibility and the Moral Sentiments, reprint edition.
[74] Linda Weber, Linda R. Weber, and Allison I. Carter. 2003. *The social construction of trust.* Springer Science & Business Media.
[75] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 195–200.
[76] Tal Zarsky. 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 (2016), 118–132.
[77] Barry J Zimmerman. 2000. Self-efficacy: An essential motive to learn. *Contemporary educational psychology* 25, 1 (2000), 82–91.