What to account for when accounting for algorithms

A systematic literature review on algorithmic accountability

Maranke Wieringa
m.a.wieringa@uu.nl
Datafied Society
Utrecht University
Utrecht, The Netherlands

ABSTRACT

As research on algorithms and their impact proliferates, so do calls for scrutiny/accountability of algorithms. A systematic review of the work that has been done in the field of 'algorithmic accountability' has so far been lacking. This contribution puts forth such a systematic review, following the PRISMA statement. 242 English articles from the period 2008 up to and including 2018 were collected and extracted from Web of Science and SCOPUS, using a recursive query design coupled with computational methods. The 242 articles were prioritized and ordered using affinity mapping, resulting in 93 'core articles' which are presented in this contribution. The recursive search strategy made it possible to look beyond the term 'algorithmic accountability'. That is, the query also included terms closely connected to the theme (e.g. ethics and AI, regulation of algorithms). This approach allows for a perspective not just from critical algorithm studies, but an interdisciplinary overview drawing on material from data studies to law, and from computer science to governance studies. To structure the material, Bovens's widely accepted definition of accountability serves as a focal point. The material is analyzed on the five points Bovens identified as integral to accountability: its arguments on (1) the actor, (2) the forum, (3) the relationship between the two, (3) the content and criteria of the account, and finally (5) the consequences which may result from the account. The review makes three contributions. First, an integration of accountability theory in the algorithmic accountability discussion. Second, a cross-sectoral overview of the that same discussion viewed in light of accountability theory which pays extra attention to accountability risks in algorithmic systems. Lastly, it provides a definition of algorithmic accountability based on accountability theory and algorithmic accountability literature.

CCS CONCEPTS

• **Social and professional topics** → **Management of computing and information systems**; **Socio-technical systems**; • **General and reference**; • **Human-centered computing** → *Collaborative and social computing*;

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372833

KEYWORDS

Algorithmic accountability, algorithmic systems, data-driven governance, accountability theory

ACM Reference Format:
Maranke Wieringa. 2020. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3351095.3372833

1 INTRODUCTION

From aviation to recruiting: it seems no sector is unaffected by the implementation of computational systems. Such computational, or 'algorithmic', systems, were once heralded as a way to remove human bias and to relieve human labor. Despite their aims, such systems were found capable of inflicting (minor to serious or even lethal) harms as well, be it intentional/unintentional. Examples of drastic situations abound. In 2019, two of Boeing's planes were presumably downed by software [71]. Volkswagen designed their cars' software to automatically cheat emission-testing [77]. Governmental systems initially designed to help now profile and discriminate the poor [63]. Amazon created a recruiting system which systematically discriminated against women, as the training data was made up of historical hiring data in which males were vastly overrepresented [47]. The effects of these systems may be intentional (e.g. Volkswagen's emission fraud), but more often are unintended side-effects, some of which may have far-reaching consequences such as the death of 346 Boeing passengers [73].

Central to such computational systems are algorithms: those sets of instructions fed to a computer to solve particular problems [70, p. 16]. As algorithms are increasingly applied within a rapidly expanding variety of fields and institutions affecting our society in crucial ways, new ways to discern and track bias, presuppositions, and prejudices built into, or resulting from algorithms are crucial. The assessment of algorithms in this matter has come to be known as 'algorithmic accountability'.

Algorithmic accountability has gained a lot of traction recently, due to the changed legislative and regulatory context of data-practice, with the implementation of the General Data Protection Regulation (GDPR), several lawsuits (e.g. A.4), and the integration with open government initiatives [65, p. 1454]. Examples of such governmental initiatives abound: the city of New York [109] installed an Automated Decisions Systems Task Force to evaluate algorithmic systems, and the Dutch Open Government Action Plan includes a segment on Open Algorithms [92]. Within civil society and academia, there are also many laudable initiatives [e.g.

3–5, 11, 50, 60, 108, 128, 138] advocating for more algorithmic accountability, yet a thorough and systematic definition of the term lacks, and it has not been systematically embedded within the existing body of work on accountability.

Nevertheless, there have been numerous works over the past decades which touch upon the theme of algorithmic accountability, albeit using different terms and stemming from different disciplines [e.g. 83, 88, 94, 116, 123]. Thus, while the term may be new, the theme certainly stands in a much older tradition of, for instance, computational accountability [e.g. 66, 112] and literate programming [88], advocating much of the same points.[1] Algorithmic accountability is thus not a new phenomenon, and accountability even less so. To avoid reinventing the wheel, we should look to these discussions and to embed algorithmic accountability firmly within accountability theory.

This contribution presents the preliminary results of a systematic review on algorithmic accountability, following the PRISMA statement [95]. 242 English articles from the period 2008 up to and including 2018 were collected and extracted from Web of Science and SCOPUS, using a recursive query design (see appendix B for an explanation of the methodology) coupled with computational methods. The material was ordered and prioritized using affinity mapping, and the 93 'core articles' which were identified as the most important will be presented in this contribution. This recursive search strategy made it possible to look beyond the term 'algorithmic accountability' and instead approach it as a theme. That is, the query also included terms closely connected to the theme (e.g. ethics and AI, regulation of algorithms). This approach, allows for an interdisciplinary perspective which appreciates the multifaceted nature of algorithmic accountability. In order the structure the material, accountability theory is used as a focal point. This review makes three contributions: 1) an integration of accountability theory in the algorithmic accountability discussion, 2) a cross-sectoral overview of that same discussion viewed in light of accountability theory which pays extra attention to accountability risks in algorithmic systems, and 3) it provides a definition of algorithmic accountability based on accountability theory and algorithmic accountability literature. In Appendix A the reader can find concrete situations which highlight some problems with accountability. These will be referred to in the corresponding sections of this paper.

2 ON ACCOUNTABILITY AND ALGORITHMIC SYSTEMS

2.0.1 Defining accountability. Making governmental conduct transparent is now viewed as 'good governance'. As such, accountability efforts can often be said to have a virtuous nature [25]. However, a side effect to such accountability efforts is the 'sunlight is the best disinfectant; electric light the most efficient policeman' [29] logic. In having to be transparent about one's work, one starts to behave better: here we see accountability used as a mechanism to facilitate better behavior [25]. Both logics can co-exist. Accountability as a term can be used in a broad and narrow sense. Typically, though, the term refers to what Bovens [24, p. 447] describes as:

> a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences.[2]

Thus an 'actor' (be they an individual, a group, or an organization) is required to explain their actions before a particular audience, the 'forum'.[3] This account is bound to particular criteria. The audience can then ask for clarifications, and additional explanations, and subsequently decides if the actor has displayed proper conduct, from which repercussions may or may not follow. What is denoted with algorithmic accountability is this kind of accountability relationship where the topic of explanation and/or justification is an algorithmic system. So what, then, is an algorithmic system?

2.0.2 Defining algorithmic systems. As noted above, algorithms are basically instructions fed to a computer [70, p. 16], They are technical constructs that are simultaneously deeply social and cultural [125]. Appreciating this 'entanglement' [13, 133] of various perspectives and enactments [125] of algorithms, this contribution sees algorithms not as solely technical objects, but rather as socio-technical systems, which are embedded in culture(s) and can be viewed, used, and approached from different perspectives (e.g. legal, technological, cultural, social). This rich set of algorithmic 'multiples' [107, cited in 125] can enhance accountability rather than limit it. The interdisciplinary systematic literature review presented in the remainder of this contribution bundles knowledge and insight from a broad range of disciplines and appreciates the entanglements and multiples that are invariably a characteristic of algorithmic system interaction.

3 WHAT TO ACCOUNT FOR WHEN ACCOUNTING FOR ALGORITHMS?

This paper draws Bovens's widely accepted definition of accountability as a relation between actor and forum is used as a focal point to structure the 93 interdisciplinary articles. The material is analyzed on the five points Bovens's identified as integral to accountability: (1) its arguments on the actor, (2) the forum, (3) the relationship between the two, (3) the content and criteria of the account, and finally (5) the consequences which may result from the account. Below, I will discuss the findings of each of these five points.

3.1 Actor

A first question would be who should be rendering the account, or who is responsible [e.g. 39, 44, 57, 89, 99, 101, 127, 142]? Aside from such a general specification, Martin [101] and Yu et al. [142] argue that one needs to specifically address two different questions. For instance, who is responsible for the harm that the system may inflict when it is working correctly [101]? Who is responsible when it is working incorrectly [142]? These questions are often not readily answerable as the organization who is using the algorithmic system need not be the developing party. In many cases organizations

[1] A clarification of the differentiation between term and theme can be found in the methodology appendix B.

[2] 'Actor' is used in the accountability-sense here, rather than in a Latourian way. Where an ANT-like actor is discussed, the term 'actant' will be used to avoid confusion.

[3] Please note that this contribution makes use of the singular they.

commission a third party to develop a system for them, which complicates the accountability relationship. When is the developer to be held accountable and when should we call the organization commissioning and using the application to the stand [57, p. 62] (cf. A.3)?

3.1.1 Levels of actors. Answering these questions is far from straightforward. In any given situation one may distinguish different actors on different *levels* of organization, making a single actor hard to pinpoint [106]. For instance, we can take the individual developing or using the system as the actor, but on higher levels one may also hold the team, the department or the organization as a whole accountable. In some instances the organization may use a system developed by a third party, which complicates the scheme even further (cf. A.3). In other words, locating the actor is often tough, and different contexts might even require different kinds of actors for the same subject.

Bovens [24] describes four types of accountability relations based on the level of the actor: *individual* accountability, *hierarchical* accountability, *collective* accountability, and *corporate* accountability. Individual accountability means that individual's conduct is held to be their own. In other words, when one is not shielded from investigation by their superiors or organization [24, p. 459]. Hierarchical accountability describes the situation in which the persons heading the organization, department or team are held accountable for that greater whole [24, p. 458]. Collective accountability rests on the idea that one can hold a member of a group of organization accountable for the whole of that organization, regardless of their function or standing [24, p. 458-459]. This kind of accountability relationship is rare in democratic contexts, as it is 'not sophisticated enough to do justice to the many differences that are important in the imputation of guilt, shame and blame' [24, p. 459]. We can speak of corporate accountability in situations where an organization as a non-human legal entity is held accountable [24, p. 458]. This is for instance the case in instances where we speak of the 'data controller' [135] or the 'developing firm' [101].

Special attention needs to be given to cases in which there is a third party who – for instance – has developed a given system for a particular organization, especially when the organization is a public institution. To illustrate, a private company may develop an fraud detection algorithm which scrutinizes people on benefits for a municipality [e.g. 131]. Martin [101] argues that in such situations, these third party organizations become a voluntary part of the decision system, making them members of the community. This willful membership creates 'an obligation to respect the norms of the community as a member' [101]. This then raises the question: how can one make sure that a third party respects the norms and values of the context in which the system will be deployed [see also 69, 124]?

3.1.2 Roles of actors. Actors can also be distinguished by their *roles*. Arguably the person drafting the specifications of the system will be put forth as an actor in different situations than a developer of the system or its user. Thus, roles can also be said to be a factor in determining the appropriate actor for particular situations. We can distinguish between three kinds of actor roles: decision makers, developers, and users.

Let us first look at decision makers, those who decide about the system, its specifications, and crucial factors. Coglianese and Lehr [41, p. 1216] note that it is important to consider 'who within an agency actually wields algorithm-specifying power'. There is much at stake in balancing which individual gets to make these decisions precisely because higher-level employees (the authors specifically discuss public administration) are more accountable to others, so they cannot be unknowledgeable about critical details of the algorithm. Here, it seems, Coglianese and Lehr refer to hierarchical accountability. They continue to argue that introducing algorithmic systems may upend work processes in a fundamental way, especially when algorithms express value judgements quantitatively, as much is lost in that translation [41, p. 1218]. Who in an organization is allowed to systematically decide how such value judgements will be structurally translated into a number? Coupled to this is the question who gets to decide when an algorithm is 'good enough' at what it is supposed to do [69]? Who, for instance, gets to decide what acceptable error rates are [41, 90] (cf. A.2)?

Developers are often seen as the responsible party for such questions as they are 'knowledgeable as to the design decisions and [are] in a unique position to inscribe the algorithm with the value-laden biases as well as roles and responsibilities of the algorithmic decision' [101]. Kraemer, Van Overveld, and Peterson [90, p. 251] are like-minded, as they note that since the developers 'cannot avoid making ethical judgments about what is good and bad, (...) it is reasonable to maintain that software designers are morally responsible for the algorithms they design'. Thus, developers implicitly or explicitly make value judgments which are woven into the algorithmic system. Here, the logic is that the choices should be left to the user as much as possible. The more those choices are withheld from users, the heavier the accountability burden for the developing entity is [90, 101].

This does imply, however, that developers and/or designers should also have the adequate sensitivity for ethical problems which may arise from the technology [130, p. 3]. Decisions about the balancing of error rates are not often part of specifications [90], which means developers have to be able to recognize and flag these ethical considerations before they can deliberate with stakeholders where needed and account for those choices. Another problem arises from what is termed the 'accountability gap' 'between the designer's control and algorithm's behavior' [106, p. 11]. Especially in learning algorithms which are vastly complex, the developer or the developing team as a whole may not control or predict the systems' behavior adequately [82, p. 374].

Special attention has to be given to the users of the system, and their engagement with it. First of all one may wonder, who is the user of the system [80, 103, 117]? Secondly, we may ask what is the intensity of human involvement? In some cases implementing algorithmic systems comes at the loss of human involvement [e.g. 41]. In general, we can distinguish between three types of systems: human-in-the-loop, human-on-the-loop, human-out-of-the-loop. This typology originally stems from AI warfare systems, but is productively applied in the context of algorithmic accountability [40, 45]. Human-in-the-loop systems can be said to augment human practice. Such systems make suggestions about possible actions, no action will be undertaken without human consent. In other words,

these are decision-guidance processes [141, p. 121]. Human-on-the-loop systems are monitored by human agents, but instead of the default being 'no, unless consent is given', this kind of system will proceed with their task unless halted by the human agent. Finally, there are human-out-of-the-loop systems where no human oversight is taking place at all. We then speak of automated decision-making processes [141, p. 121]. Arguably, these different kinds of involvement have consequences for the accounts which can be rendered by the user-as-actor. Thus, one aspect of an account of algorithms should be the measure of human involvement [37, 40, 41, 46, 62, 89].

3.2 Forum

As one sets out to account for their practice, it is important to consider to whom that account is directed [33, 86, 103, 135]. Kemper and Kolkman [86] argue that one cannot give account without the audience understanding the subject matter and being able to engage with the material in a critical way. Their argument for the 'critical audience' shows parallels with Bovens's articulation of accountability in which 'the forum can pose questions and pass judgement' [24, p. 450].

What shape can this critical audience take, then? The EU's [64] General Data Protection Regulation (GDPR), hailed partly for its 'right to explanation', may point towards the individual citizen as the forum in the context of algorithmic accountability [135, p. 213-214]. In other cases, one may need to give account to one's peers, or the organization accounts to an auditor [32, p. 318]. Different fora can be interwoven, but each requires different kinds of explanations and justifications [cf. 21, 22, 142].

Bovens [24] describes five kinds of accountability relations based on the type of forum: *political* accountability (e.g. ministerial responsibility; cf. A.4), *legal* accountability (e.g. judges; cf. A.4), *administrative* accountability (e.g. auditors inspecting a system), *professional* accountability (e.g. insight by peers; cf. A.3), and *social* accountability (e.g. civil society).

Political accountability can be said to be the inverse and direct consequence of delegation from a political representative to civil servants [24, p. 455]. As tasks are delegated, the civil servant has to account for their conduct to their political superior.

What has changed is that not only do politicians delegate to civil servants now, but civil servants themselves start to delegate to and/or are replaced by algorithmic systems. This change is one that has been identified before by Bovens and Zouridis [27] in connection to the discretionary power of civil servants. Bovens and Zouridis note that civil servants' discretion can be heavily curtailed by ICT systems within the government. Building on Lipsky's [96] conception of the street-level bureaucrat, they make a distinction between street-level bureaucracy, screen-level bureaucracy, and system-level bureaucracy. Each of these types of bureaucracy allow for different measures of discretionary power of civil servants. Whereas street-level bureaucrats have a great measure of discretion, screen-level bureaucrats' discretionary power is much more restricted. System-level bureaucracy allows for little to no discretionary power, as the system has replaced the civil servant entirely.

The different forms of bureaucracy are coupled to the way in which systems are playing a role within work processes. As Bovens and Zouridis note, the more decisive the system's outcome is, the less discretion the user has. Delegation to systems is thus not a neutral process, but one that has great consequences for the way in which cases are dealt with, and border cases especially. Delegation to systems is important to consider for two other reasons as well. First, following Bovens' [24] logic of delegation and/or accountability it would make sense to start to hold the algorithmic system accountable.

There are many efforts to make algorithms explainable and intelligible. Guidotti et al. [72], for instance, note that we can speak of four different types of efforts to make a 'wicked' algorithm intelligible [10]. These four approaches also correspond to efforts discussed in the field of explainable AI (XAI) [e.g. 2]:

(1) Explaining the model, or: 'global' interpretability;
(2) Explaining the outcome, or: 'local' interpretability;
(3) Inspecting the black box;
(4) Creating a 'transparent box'.

An explanation of the model is an account of the global logic of the system, whereas an explanation of the outcome is a local and, in the case of personalized decisions, personal. Inspecting the black box can take many shapes, such as reconstructing how the black box works internally, and visualizing the results (cf. A.1). In other cases auditors may be used to scrutinize the system [32, p. 318]. Another approach would be to construct a 'transparent box' system which does not use opaque or implicit predictors, but rather explicit and visible ones.

Such technical transparency of the working of the system can be helpful, but in itself should be considered insufficient for the present discussion, as accountability > transparency. The transparent workings of a system do not tell you why this system was deemed 'good enough' at decision making, or why it was deemed desirable to begin with [102]. Nor does it tell us anything about its specifications or functions, nor who decided on these, nor why [41, p. 1177]. Whereas transparency is thus passive (i.e. 'see for yourself how it works'), accountability requires a more active and involved stance (i.e. 'let me tell you how it works, and why').

Second, whereas, in the context of the government, civil servants have the flexibility to subtly shift the execution of their tasks in light of the present political context, systems do not have such sensitivity. Often, such systems are not updated to the contemporary political context, and thus 'lock in values' for the duration of their lifecycle (cf. A.1). Accounts on algorithms are thus key as algorithmic systems are both 'instruments and outcome of governance' [84, following 85]. They are thus tools to implement particular governance strategies, but are also themselves a form of governance. Thus, accountability is crucial if we wish to avoid governance effects through obsolete values/choices, embedded in algorithmic systems.

Legal accountability is usually 'based on specific responsibilities, formally or legally conferred upon authorities' [24, p. 456]. Much of the actions systems will undertake are not up for deliberation, as they are enshrined in law [62, p. 413]. There are thus already laws and regulations which apply to systems and can be leveraged to ensure compliance.

However, as Coglianese and Lehr [41, p. 1188] note, laws do not prescribe all aspects of algorithmic systems. For instance, there is no one set acceptable level of error. Rather, the acceptability strongly depends on the context and use of the system [90, 102] (cf. A.2). There is thus also a matter of discretion on the part of the system designers on how one operates within the gaps of the judicial code, in these cases 'ethical guidance' needs to come from the human developer, decision maker, or user [62, p.416-417].

This does raise some questions with regards to the ethical sensitivity of these human agents. As it stands, technical experts may not be adequately aware of the laws and legal system which they operate in [46]. On the side of the legal system, there may also be insufficient capacity to understand algorithmic systems. We see this in cases where lawyers and expert witnesses must be able to inform their evidence with the working and design of the system [33]. Algorithmic systems thus require new kinds of expertise from lawyers, judges, and legal practitioners in general [e.g. 75, p. 75], as they need to be able to assess the sometimes conflicting laws and public values, within those systems, which themselves can be vastly complicated and rely on a large amount of connected data sources. Meaningful insight in this interwoven socio-technical system [31] is needed to decide whether these values are properly balanced and adequately accounted for. Yet it can be particularly hard to decide on what may provide meaningful insight and what will result in 'opacity through transparency' [113, 118].

While the data gathering phase is quite well regulated, the analysis and use phases of data are underregulated [32], leaving judges to fend for themselves. What might prove to be crucial is, however, the socio-technical aspect of the system. For as has been discussed above, algorithmic systems 'do not set their own objective functions nor are they completely outside human control. An algorithm, by its very definition, must have its parameters and uses specified by humans' [41, p. 1177]. Eventually, someone has made choices about the system, and mapping these pivotal moments might help in resolving a transparency overload of information.

These complications of expertise in the legal forum deserve special attention as it has a facilitating function. The legal framework provides many other fora (e.g. civil society) with the means to enforce accountability, for instance through freedom of information (FOI) requests [65]. Though as Fink [65] notes, FOI requests often have a limited use due to the reserved exemptions. Nevertheless, jurisprudence and regulation are often enablers to build cases and enforce accountability [22]. As such the legal system is particularly important for other accountability arrangements as well.

Administrative accountability refers to 'a wide range of quasi-legal forums, exercising independent and external administrative and financial supervision and control' [24, p. 456]. Examples of such administrative accountability fora are safety certificators, accident investigators [33], auditors [106, p. 13], and regulators [135]. Domain-specific authorities are another form of administrative fora.

Many authors note that adequate administrative authorities are lacking, and argue these should be instituted [e.g. 40, 41, 55, 57, 127, 135]. The precise shape these administrative authorities should take is largely left underdeveloped, however.

Professional accountability deals with those kinds of accountability relations between a professional and their peer group [24, p. 456-457]. 'Peer group' is here interpreted in a loose fashion to denote fora within one's organization [cf. 134]. Acknowledging internal fora is important, as accountability practices need to be entrenched in the organization's structure in order for external accountability to be viable [134].

Professional fora outside of the organization may be associations of particular disciplines. Here, an accountability relationship may comprise the adherence to the many guidelines and standards articulated by such organizations [e.g. 1, 38, 79]. Some of these standards are norms or best practices to that function as accountability mechanisms, others are moral imperatives. Brennan-Marquez [31, p. 1297], for example, notes how explanatory standards 'create incentives for institutional actors (…) to understand the tools they employ'. Diakopoulos [51, p. 58] for instance illustrates the moral ethos which some of these guidelines advocate quite apt with the ACM Code of Ethics [1] for software engineering:

> First and foremost is that software engineers should act in the public interest: to be accountable and responsible for their work, to moderate private interests with public good, to ensure safety and privacy, to avoid deception, and to consider the disadvantaged. The general moral imperatives of ACM include "avoid harm to others," "be fair and take action not to discriminate," and "respect the privacy of others." Let that sink in.

Here, we see how such guidelines do not lean on an idea of accountability as a mechanism, but rather on accountability as a virtue.

Another type of professional accountability may deal with the modularity and ecological nature of algorithms. Algorithms never exist in a void, but rather are dependent on one another [97]. Thus, there is also needs to be an accountability between developers of dependent systems. Torresen [130, p. 4] notes, for instance, the importance of control mechanisms between systems – which thus require a coordination between different teams of developers/decision makers/users.

Lastly, there is the accountability among different kinds of professionals. To illustrate, the developer of a system can also be said to be accountable to the user and/or decision maker about their embedded value judgements, for instance [cf. 90].

Finally, Bovens [24, p. 457] discusses social accountability. Social accountability can take the form of 'more direct accountability relations between public agencies, on the one hand, and clients, citizens and civil society, on the other hand'. Fora connected to this kind of accountability are for instance NGO's, and interest groups, but also individual citizens. In other words, this kind of accountability relationship deals with the wider society [33]. Such accountability is key as humans (i.e. developers, users of the system) do not only shape the algorithm, but the algorithm also shapes humans (developers, users, subjects) and society [84, p. 252]. Rahwan [120] notes in this light that we perhaps should not just have a human-in-the-loop, but also enforce a social contract in which values of stakeholders are negotiated. In this way, he argues, we can keep 'society-in-the-loop', and safeguard public values. Similarly, Fink [65, p. 1454] argues that algorithms should be made inspectable for

a 'broad range of people', and Janssen and Kuk [82, p. 372] note that citizens should be able to scrutinize the government's algorithms.

3.3 The accountability relationship

Accountability relationships between actor and forum can, as we have seen, come in several different shapes, levels, extents of disclosure and discussion, and differing severities of consequences. Yet, all these relationships follow a particular rhythm as they all go through three phases [30, p. 960-961]. The first is the information phase, in which the actor gives information to the forum. The second phase is the deliberation and discussion of the forum, and the questions asked to the actor. The final phase concerns the consequences imposed on the actor by the forum. These phases can be mapped on a spectrum of quantity/intensity. This map, the 'accountability cube', is a three-dimensional representation of the three consecutive phases of the accountability arrangement: information-giving, discussion, and the imposing of consequences. Each of the phases can be 'measured' separately, giving little information does not necessarily entail little discussion amongst the forum, for instance. As such, it makes sense to reflect on each of the three 'scales' apart from one another. The cube serves as a tool to assess accountability relationships, and to empirically identify accountability deficits and overloads [30, p. 960-961].

What the cube does not specify, is the shared understanding and perspective on which underlies the accountability relationship. After all, accountability efforts adequate in one situation may be insufficient in others. Bovens, Schillemans, and 't Hart [26] distinguish three normative perspectives on accountability: a democratic perspective, a constitutional perspective, and a learning perspective.

The democratic perspective departs from the idea that accountability 'controls and legitimizes government actions by linking them effectively to the "democratic chain of delegation"' [26, p. 231]. The success of accountability, viewed in this light, is measured in the degree to which accountability helps to assess the executive branch, and in how it works as a mechanism to enforce better behavior. A constitutional perspective argues that accountability plays a crucial rule in withstanding 'the ever-present power concentration and abuse of powers in the executive branch' [26, p. 231]. Successful accountability in such a perspective prevents the abuse of one's executive abilities. This perspective is thus concerned with preventing corruption and safeguarding the integrity of the executive branch of government. A learning perspective on accountability sees it as a way to provide 'public office-holders and agencies with feedback-based inducements to increase their effectiveness and efficiency' [26, p. 232]. Here, the evaluation standard of accountability concerns the degree in which an accountability arrangement successfully stimulates a focus on societally desirable outcomes.

3.4 The account and its criteria

An account can come in many forms, and at many moments in the algorithm's lifecycle. Before we ask what that account should entail, let us first dwell on when in the lifecycle, and at which point in its deployment, the account of the algorithm is, could or should be rendered. Kroll et al. [91] note that, traditionally, there are two approaches to such evaluations: ex ante (before the fact) and ex post (after the fact). With algorithms though, they and several others posit that accountability should be kept in mind throughout the whole design process [51, 91, 111]. Below, these standpoints and arguments are discussed in more detail.

Several scholars point to ex ante evaluations such as impact assessments [e.g. 40, 135] or simulations of behavior [12]. Those evaluations are always limited, as one cannot foresee the entire process of an algorithm's deployment [82, 135]. Nevertheless, the importance of an accountability relationship arguably also depends on the extent to which it impacts society, and individuals, and the role of the algorithm in that decision [74, cited in 75]. Martin [101] notes that we need to weigh the role of an algorithmic decision in the decision-making process, and the impact of the final decision on individuals and the wider society. Weighing these factors might provide some guidance on how thorough and extensive future accounts need to be.

There are also those who explicitly warn against rendering technology accountable before the fact, as 'we risk attributing certainty and responsibility for such a future path to the algorithm' [110, p. 52]. Some argue that we can only meaningfully account for algorithms after the fact, because of the nature of big data research which tends to search for new applications for the same data [135].

Finally, there are those who argue that we need to consider algorithms not just before the fact, and/or after the fact. Instead, one needs to consider the entire process: the design, the implementation and the evaluation [e.g. 51, 91, 110]. Neyland [110] most notably, contributed to the design of an ethical surveillance system and employed anthropological and ethnomethodological techniques to give account of the system's development process. As he illustrates [110, p. 68]:

> Accountability was not accomplished in a single moment, by a single person, but instead was distributed among project members and the ethics board and across ongoing activities, with questions taken back to the project team between meetings and even to be carried forward into future projects after the final ethics board meeting.

An account of algorithms, like the design and the execution of the system, unfolds over time. As Kate Crawford [42, p. 79] argues, a sole focus on the outcome of the algorithm 'forecloses more complex readings of the political spaces in which algorithms function, are produced, and modified'. Pivotal moments such as the choice for a particular algorithm [e.g. 56, p. 549], and other design decisions [101] such as the weighting of factors [40, p. 17], or balancing of 'fairness' [59] deeply influence the system. Such decisions are generally informed by tests with different implementations of the system, each of these versions inform the final implementation in one way or another. The building of an algorithmic system is thus an incremental process of assemblage [cf. 82], that cannot be equated to a final product at any point, which is why disclosing the tests done during the process might be very informative [41, p. 1212]. This contribution sides with the later view, in which algorithms are not something that can be assessed in a single moment, but that assessment should follow the system's lifecycle.

As we saw earlier, current efforts of making algorithmic systems explainable, most notably Explainable AI (XAI), tend to focus on a

technical transparency of specific aspects of the algorithmic system. The primary goal of these approaches seems to be the transparency of the system rather than justification of the system [cf. 118]. This is where we touch upon the socio-technical aspect of algorithms. This is where the field of XAI and explainable algorithms tends to fall short. As such, this contribution puts forth a fifth approach, which is also practiced/advocated by others such as Neyland [110] and Gasser and Almeida [67] which combines much of these initial four strategies, but also affords explanations on the socio-technical nature of the system and respects the temporal unfolding of an algorithmic system. Such a socio-technical account can encompass, amongst others, the algorithmic system's reason for existence, the context of the development, the effects of the system. Yet the socio-technical account should not be seen as a checklist of everything that needs to be addressed, but rather as a modular frame which can help identify and ask the questions crucial in particular contexts.

Instead of it being a dichotomous either/or, a modular account allows for more attention to the crucial considerations, and affords paying less attention to less relevant ones. Rigorous assessment of every algorithm is unworkable, and as politicians have rightly pointed out [e.g. 49], it would be too costly. Thus, engaging with a modular accountability framework for algorithms could help balance on the one hand the costs, and on the other the public's right to information and explanation. For instance, as a starting point, one could pay the most attention to systems which substantially impact individuals [22, 45].

Below I will highlight what aspects of a modular account of the socio-technical algorithm could be. I divide this account in ex ante, in medias res, and ex post considerations. The different considerations play at different moments in the software development life cycle (SDLC). The SDLC is made up out of six stages: planning, analysis, design, implementation, testing/integration, and maintenance [78]. Planning is about articulating specifications and user needs, identifying the desirability of the software, and creating a strategy for development. Analysis is about translating the specifications and goals of the project to functionalities, and identifying and tackling hindrances to a successful software implementation. Together, these two stages form the ex ante considerations of an algorithmic system, for it is only after these stages that that which we tend to understand as 'software development' (i.e. coding, implementation, testing) comes into the picture, as part of the in medias res considerations. Here we touch upon the SDLC stages of design, implementation and testing/integration. Design is about creating the architecture for the application. Implementation is where the programming of the software happens, generally this happens in a modular way; that is, programmers/teams each work on separate aspects of the system. Testing and integration is where the separately produced aspects of the product are connected, that is integrated. The integrated whole is subsequently tested 'in vitro' [134] for errors, bugs, and other unforeseen issues. Finally, there is the maintenance stage, where the product is deployed, the software needs to maintained and the 'in vivo' [134] bugs needs to be resolved. This stage also requires ongoing evaluations of the product's quality and relevance. It would be tempting to locate the ex post considerations solely with this last maintenance stage, but there is in fact much more to it. Many important decisions, for instance relating to disclosure, are arrived upon earlier. Moreover, the system may inform the planning phase of other, to be developed, systems. An example can be a system which uses decisions of other systems for its own processes. Similarly, ex ante/in medias res considerations are also less clearly demarcated than an initial mapping would suggest. The reason for much of this 'bleeding over' is that the SDLC is non-linear, meaning that sometimes one needs to return to earlier stages of the life cycle. If, for instance, tests expose a mismatch between the context and the conceived product, one may have to revisit the plan, analysis and/or design.

A distinction in life cycle stages, as any distinction, is thus always an artificial one, as is the separation between ex ante/in medias res/ex post considerations. Nevertheless, these demarcations help to identify and type what accounts are needed at what stage of the design of the algorithmic socio-technical system.

3.4.1 Ex ante considerations. If we consider the account to be something that tells of and justifies one's conduct, we see a great parallel with storytelling. What is different between storytelling and rendering an account is that the audience, or forum, often has the ability to impose sanctions. In fact, as any parent (or child, for that matter) may tell you, there is a thin line between 'telling what has happened' and 'accounting for what has happened'. Suddenly a story may have consequences after all, once a child innocently tells you about one of their dangerous endeavors of the day.

Stories, like accounts, need ingredients in order for them to be sensible. At their basis, they require a who, what, where, when, and why. These interrogative words, dubbed the 5 W's, are needed for stories because they situate actions (what) concretely (where, when, who), and specify the underlying logic (why). Because of their situating capacities, these words can prove beneficial as formal focal points for accounts as well.

In the following, I will use these W's to structure the account of ex ante considerations around algorithmic systems, making one addition to the list: whom it affects. As we have seen, the socio-technical algorithmic system is complex. There are a lot of groups coming together around the system: developers, users, decision makers, but these systems affect people as well, for instance citizens/consumers. Not all of these groups have a similar amount of power, as was discussed earlier in this contribution. As such, it is beneficial to make a distinction between who is creating/using the system and whom the system affects.

Who is developing and using the system matters, for these persons influence the system. A crucial aspect of this first element concerns the question whose values are informing the system. McGrath and Gupta [102] note that one of the key distinctions between humans and algorithmic systems is that while humans are able to negotiate conflicting values or rules, algorithmic systems need a prioritization of those values. One thus needs to account how those values have been balanced [31]. There are many ways in which one can think about the balancing of those values, such as crowdsourcing, or people's councils [103]. Yet this does not entirely solve the problem, for as Baum [15] asks: whose considerations and norms and values are included in the design of the system and whose are left out? People's councils or crowdsourced initiatives can strive to but never are a true cross-section of society. Keeping this in mind is important for as Kraemer, Van Overveld, and Peterson [90, p. 251] note, 'two persons or more who accept different value-judgements

may have a rational reason to design [the algorithm] differently'. In other words, even in collaboration with stakeholders such as civil society or people's councils, we still need an account of the preferences and choices which inform the system's design. Moreover, deciding on such a strategy is itself an important design choice. This implies two things. First, there will inevitably be friction between values/value judgements among those involved. Second, the process in which the decision was made to prioritize one value/value judgement over other possibilities needs to be accounted for; that is, we must account for the development history of the entire assemblage [7, p. 109].

Connected to this is the question where the system is being developed and deployed [42, 51, 54, 86, 104, 134]. This is even more crucial when a third party is developing the system, as they will need to respect the norms and values of the context in which the system will eventually be deployed [69, 101, 124]. Moreover, joining forces with a third party also limits the options of main organization to render account, as commercial interests may prevent certain acts of transparency and/or justification [118].

This leads us to a second question of what it is that the organization/organizations set out to create precisely. This part of the account should explain what the system is intended to do [34, 72]. This is relatively simple, as many projects already have specifications of a given system before they are developed.

A third question is why this system is needed, and why it should take the proposed shape. This concerns the system's raison d'être, that is, why the system is needed in the first place. This includes a reflection on what it will change about the existing situation [36, 80], as well as the system's envisioned place within work processes and the organization. We must also ask why does it take this form? What, if any, alternative solutions/systems were considered [69], why were those rejected, and what makes the arrived upon option the most socially desirable [15, 37, 41, 117, 135]? In other words, one needs not only justify the development of the system, but also its implementation [91] in light of its socio-technical situatedness.

A fourth element inquires when the system is developed, maintained, and terminated. As algorithmic systems are situated in time, they need to be periodically assessed for their contextual fitness. It is important to sketch the temporal frame in which the system was originally conceived and justified, which provides a touchstone for future evaluations. These evaluation moments serve to assess three things [32, p. 318]. First the need for the system needs to be assessed. As the context of the system changes, the system may no longer be necessary. Secondly, the working of the system needs to evaluated. Here the specifications and benchmarks, if any, can serve as a touchstone, to assess the effectiveness of the system. Third, they allow for checking whether the initial assumptions and conditions, from which the project departed, still hold true.

The latter point is often glossed over in system evaluation, but is crucial. Without such evaluation, there is a risk that 'a piece of software locks in a particular interpretation of law or policy for the duration of its use, and, especially in government contexts, provisions to update the software code may not be made' [91, p. 701] (cf. A.1). Regular evaluations are thus a necessity to avoid legal and value lock-in. As Just and Latzer [84, p. 254] note, algorithmic systems have a similar effect on society as laws and contracts do. As such, they too should be open for periodical scrutiny.

Finally, we need to consider whom it affects. What groups, people, and/or situations will the system affect and in what way [84, 103, 104, 127, 135]? Accounting for this element can be done by making impact assessments [40, 135]. There are many types of impact assessments, which might or might not be useful depending on the situation, such as, but not limited to, the Privacy Impact Assessment (PIA), the Artificial Intelligence Impact Assessment (AIIA), and the Data Protection Impact Assessment (DPIA). Each of these focuses on specific aspects of the system and its implications. Such impact assessments allow for evaluating proportionality of the system [32, p. 318].

3.4.2 In medias res considerations. Just as stories and accounts are build up out of the W's, they also address how things have happened. This is where we touch upon the in medias res considerations of design, implementation, and testing and integration. First of all, we may ask how does the system work. Is it a simple decision tree, a rule-based model, or is it technically more complex and does it use, for instance, machine learning?

Coupled to this, we must ask how we can be sure that the system is accurate and fair? Especially in cases where machine learning and artificial intelligence in general are used, we should be attentive to how they are employed. Machine learning tries to learn from historical data, to subsequently be applied in new circumstances. As such circumstances change, and historical data may be imbued with biases, systemic or otherwise, we need to ask whether the historical training data chosen is a fair and appropriate reference point for this decision [56, 86, 91, 101, 127].

Machine learning comes in different flavors, but at its two extremes are supervised learning and unsupervised learning. In supervised learning, the algorithm is trained to recognize particular features or patterns by using labeled data. While this increases explainability to a large extent, as one knows what categories have been fed [56], surprises may occur [cf. 121]. In unsupervised systems the system is fed unlabeled data and needs to figure out patterns for themselves. For both types it is crucial to assess how accurate the system is [117].

A lot of this boils down to rendering account of how the system has been tested, and what the subsequent results were [91]. Moreover, are those results used to change anything about the system, and if so, what is changed and why [41]?

In line with fairness, we may ask how membership of protected classes is kept out of the system? In many contexts, the consideration of race, ethnicity, sexuality, gender, religion, and others is not allowed as these are protected under non-discrimination principles. Yet, with machine learning, how can we make sure that the system does not explicitly, but more importantly, implicitly consider such protected classes [56, 91], for instance by using proxies which correspond to such protected classes (e.g. zip-codes as a stand-in for race and ethnicity, particularly in the USA).

In learning systems, one may also wonder whether the training is continuous or whether the model is merely trained before the fact and then deployed. In the former case, it is appropriate to ask how you will make sure that the learning algorithm will stay fair [91] and accurate.

We may also wonder how the system settles on a decision. There are different types of decision-making processes, namely those

based on prioritization, classification, association, filtering [51, 54], randomization [91], and prediction [40]. Each of these requires answering different kinds of questions.

We may for instance ask how the system was designed to be used. Here we touch upon the agency of the user. That is: are decisions made by the system automatic, or do they guide and/or support human decision-making [141]? In other words, how much discretion is the user allowed to have [40, 101, 124, 127]? We may also wonder how much information the user is given as to the decision/workings of the system [40, 101], and whether or not they can dispute that decision [40, 56, 101, 124].

Taking into account the iterative nature of software development, we may also question how the system has changed over time, why it changed, under what circumstances, and how extensive those changes were [36, 43, 110, 136]. Most importantly, one should ask whether or not these changes require a revisit of earlier considerations about the ethics/accountability around the system [80].

3.4.3 Ex post considerations. Finally, we can look back to what has been done, and look forward to new situations which may arise. Thus, we consider questions of whence we came, and whither we go. It is not uncommon for systems to subtly divert from their original specification. As we have seen, algorithms are not fixed objects, as they are often updated, tweaked and/or their behavior changes through machine learning. Thus, one needs to check whether the system still confirms to the initial specifications [91], and the values which underlie the project [37, 117].

One may also wonder what they can explain/disclose about how the system's decision came to be to, the data subject or other kinds of fora [37, 44, 53, 54, 135]. This also touches upon legal considerations where personal data is concerned. The GDPR includes, for instance, a 'right to explanation' in cases where automated decision-making is involved.

3.5 Consequences

Finally, the accountability relationship requires that the forum can impose consequences. Consequences can come in many different forms, which is closely connected to the kind of obligation the actor has to the forum. Bovens [24, p. 460] notes three different kinds of accountability, based on the nature of the power relation which exists between the actor and the forum: vertical accountability, horizontal accountability, and diagonal accountability.

Consequences are made most tangible when there is a vertical accountability relationship between actor and forum. Here, 'the forum formally wields power over the actor' [24, p. 460]. As one may suspect, this is the case in many instances of political and legal accountability, but also in disciplinary hearings, for instance, which are a form of professional accountability.

Cath et al. [37] note that governments have key role in creating new policies (specifically with regards to artificial intelligence). However, as Metzinger [105] warns such policy initiatives may end up being co-opted by the industry, as has happened to the High Level Expert Group for AI Ethics, he argues. He warns that, ultimately, such initiatives may end up as a toothless version of the thing they set out to be. It is precisely those teeth which allow the neccessary enforceability in a vertical power relationship.

What shape could consequences take in vertical accountability? Wagner [136] describes cases in which the automated output is subsequently redacted by humans, so as to comply with user/legal requests. Here he makes a distinction between the first order rules embedded in code, and the second order rules which are the manual changes to the output. Here, the actor is required to revise their course of actions so as to comply with the ruling of the forum.

On the other end of the spectrum stands horizontal accountability. This accountability relation based more on a moral imperative, instead of a formal one. One way in which such morally informed horizontal accountability is expressed is through self-regulation of organizations. According to Saurwein, Just, and Latzer [124, p. 39] such 'self-organization measures include company principles and standards that reflect the public interest, internal quality assessment in relation to certain risks and ombudsman schemes to deal with complaints'.

An example of self-regulation is for instance the Partnership on AI [115], which was founded in 2016 by Amazon, Facebook, Google, DeepMind, Microsoft and IBM, and aims to establish best practices, increase algorithmic literacy, and to highlight AI applications for 'socially beneficial purposes' [114]. While commendable, a risk with this kind of initiatives is that they are a form of 'ethics washing' [e.g. 87, 105, 129, 137], or rather 'virtue-washing'. As Pasquale and Citron [40, p. 22] note self-regulation does not address the organization's first obligation to efficiency rather than public values and human rights [see also 46]. Saurwein, Just and Latzer [124, p. 39] also argue that such self-regulation may serve to 'increase reputation or to avoid reputation loss'. There is thus a risk that an organization is concerned with the display of good behavior for ulterior motives, rather than with responsible behavior itself. Much of these risks lie in the nature of the obligation to the forum. As the forum cannot enforce accountability, little to no consequences can be imposed (aside from public outrage). Thus, one risks entering a slippery slope of non-committal ethics initiatives which cannot be enforced. On the other hand, other scholars such as Doneda and Almeida [57, p. 62] see self-regulation as something that could work effectively, provided that organizations and the industry as a whole implement administrative bodies which can safeguard public values. As noted above, such virtue-washing may be a way out of a vertical accountability arrangement in favor of a horizontal power relation. Diagonal accountability is an in-between form of accountability where the forum has no or little formal power over the actor. It is quite often found in administrative accountability settings, for instance in relation to ombudsmen or auditors [24, p. 460]. As was mentioned earlier when discussing administrative accountability, there is a great call for more such accountability, but little practical suggest as to how to design it.

3.6 Accountability risks

The literature review identified several accountability risks which we will enumerate below:

(1) Actor-related accountability risks;
   - The problem of many hands;
   - Interdisciplinary miscommunication;
   - Unfamiliarity with the domain;
(2) Forum-related accountability risks;

- Problem of many eyes;
- Delegation to systems may impact political accountability;
- Dependencies and power relations between different roles;
- Right to inform legal procedures in black-box systems;
- Insufficient administrative oversight bodies;
- Miscommunication between disciplines;
- Virtue-washing;

(3) Relationship-related accountability risks;
- Perspective/perspectives not defined;
- Perspectives merging or blurring;
- Deficits and overloads;

(4) Account-related accountability risks;
- Phases may not be easy to differentiate;
- Which actor should account for what;
- Incompleteness;

(5) Consequences-related accountability risks;
- Virtue-washing;
- Little means to enforce.

Some of these are general accountability risks (e.g. problem of many hands). In such cases, it would do well to turn to accountability theory and learn other domains which have tackled such problems. Other risks are 'medium-specific'. Both could be avenues of further research within FAT*, yet they require a different kind of interdisciplinarity.

4 INVESTIGATING ALGORITHMIC ACCOUNTABILITY

What this systematic literature review demonstrates is that we need to move to an accountability relationship not just of the use, the design, the implementation, or the consequences of algorithmic systems, but to consider the entirety of that socio-technical process. While the term 'algorithmic accountability' is inherently vague, as it leaves a lot room for specification about the accountability relationship, it can be specified as follows:

> Algorithmic accountability concerns a networked account for a socio-technical algorithmic system, following the various stages of the system's lifecycle. In this accountability relationship, multiple actors (e.g. decision makers, developers, users) have the obligation to explain and justify their use, design, and/or decisions of/concerning the system and the subsequent effects of that conduct. As different kinds of actors are in play during the life of the system, they may be held to account by various types of fora (e.g. internal/external to the organization, formal/informal), either for particular aspects of the system (i.e. a modular account) or for the entirety of the system (i.e. an integral account). Such fora must be able to pose questions and pass judgement, after which one or several actors may face consequences. The relationship(s) between forum/fora and actor(s) departs from a particular perspective on accountability.

First, the algorithmic accountability relationship is 'networked' and accountability is thus dispersed among many different actors [106]. It is thus key to concretely specify the actors, their role, level and the part of the system for which they are responsible. Second, we see that different fora come into play, instead of the traditional singular forum [cf. 24]. However, one forum requires a different account than another, thus it is necessary to clearly delineate to what fora one caters, and what each of these fora needs. Third, the account itself can be divided into three types of considerations, which can also be mapped to the SDLC and the relevant actors: ex ante, in medias res, and ex post considerations. This also touches upon the criteria of the account, for instance of when to explain/justify what portion of the system. Fourth, there are consequences which may be imposed on the actor by the forum. Here we can distinguish the power relation between the actants [6], and the amount of consequences imposed. Fifth, and finally, it requires active consideration of the perspective on the accountability arrangement, which may in some cases overlap. Making clear what the main perspective is of the accountability arrangement is thus important, as it helps to identify what needs to be accounted for in the algorithmic system. While the latter elements of consequences and perspectives are rather general, we must not lose sight of their importance, else we fall into the trap of virtue-washing or ill-defined expectations about the system's accountability requirements.

This definition, grounded in accountability theory [24], envelops the work that has been done in the past [e.g. 8, 110], and invites future research into the complex and interwoven networked accountability-relations surrounding algorithmic systems. At its core, this definition identifies five elements needed for the accountability arrangement: actor(s), forum/fora, perspective, account, and consequences.

Each of these elements can have a high or low intensity in the accountability relationship. The actor is scaled on how well the actor is specified (unspecified <> specified), the forum on the intensity of the discussion (non-intensive <> intensive), the perspective on its clarity (undefined <> defined), the account on its comprehensiveness (little <> much), and, finally, the amount of consequences (few <> many). However, there is a Goldilocks-effect to the accountability arrangement. Too little of an aspect risks a deficit. Too much of an aspect risks an overload. If we aim to establish an effective accountability arrangement, we will have to balance each aspect's scale (e.g. making sure the account is comprehensive enough, but not overly detailed) so that workable accountability is achieved.

Though a cross-disciplinary systematic literature review is necessarily an abstraction, there are two important take-aways that are worth mentioning. First, accountability theory is sparsely referred to, and the field would do well to take note of accountability theory which originates in governance studies. Second, as this is an issue that affects many disciplines and practices interdisciplinary engagement is a prerequisite. Neither law, critical data/algorithm studies, governance studies, data science, and the various domains in which these algorithmic systems are applied, can tackle these questions alone. As algorithmic systems are 'multiple' so should our efforts to hold them accountable be. This contribution furthers these goals. As this is a cross-sectoral overview, further research is needed to ground accountability theory and an interdisciplinary perspective on the algorithmic 'multiple' in the respective domains in which algorithmic accountability is required. Moreover, a promising avenue of research could be a mapping of the discrepancies between the fields' perspectives on the matter.

REFERENCES

[1] ACM. 2018. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Optimity advisors and Future of privacy forum. 2018. *algo:aware: Raising awareness on algorithms.* Technical Report. Washington.

[4] AI Now. 2018. *Algorithmic Accountability Policy Toolkit.* Technical Report. AI Now / New York University, New York.

[5] AI Now. 2018. *Litigating algorithms: challenging government use of algorithmic decision systems.* Technical Report. AI Now/New York University, New York.

[6] Madeleine Akrich and Bruno Latour. 1992. A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. In *Shaping technology / building society: studies in sociotechnical change*, Wiebe E. Bijker and John Law (Eds.). MIT Press, Cambridge/London, Chapter 9, 259–264.

[7] Mike Ananny. 2015. Toward an Ethics of Algorithms. *Science, Technology, & Human Values* 41, 1 (2015), 93–117. https://doi.org/10.1177/0162243915606523

[8] Mike Ananny. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science Technology and Human Values* 41, 1 (2016), 93–117. https://doi.org/10.1177/0162243915606523 arXiv:arXiv:1011.1669v3

[9] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645

[10] Leighton Andrews. 2018. Public administration, public leadership and the construction of public value in the age of the algorithm and 'big data'. *Public Administration* (2018). https://doi.org/10.1111/padm.12534

[11] Theo Araujo, Claes De Vreese, Natali Helberger, Sanne Kruikemeier, Julia Van Weert, Bol, Nadine, Daniel Oberski, Mykola Pechenizkiy, Gabi Schaap, and Linnet Taylor. 2018. *Automated Decision-Making Fairness in an AI-driven world: Public perceptions, hopes, and concerns.* Technical Report. University of Amsterdam, Amsterdam.

[12] Thomas Arnold and Matthias Scheutz. 2018. The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology* 20, 1 (2018), 59–69. https://doi.org/10.1007/s10676-018-9447-7

[13] Karen Barad. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.* Duke University Press, Durham/London.

[14] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the Third International ICWSM Conference.* http://www.aaai.org/ocs/index.php/ICWSM/09/paper/download/154/1009

[15] Seth D. Baum. 2017. Social choice ethics in artificial intelligence. *AI and Society* (2017), 1–12. https://doi.org/10.1007/s00146-017-0760-1

[16] Beall's List. 2018. Beall's list of Predatory Journals and Publishers. https://beallslist.weebly.com/

[17] Lyria Bennett Moses and Janet Chan. 2018. Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing and Society* 28, 7 (2018), 806–822. https://doi.org/10.1080/10439463.2016.1253695

[18] Bij Voorbaat Verdacht. 2018. Wat is SyRI? https://bijvoorbaatverdacht.nl/wat-is-syri/

[19] Bij Voorbaat Verdacht. 2019. Missie. https://bijvoorbaatverdacht.nl/missie/

[20] Reuben Binns. 2017. Algorithmic Accountability and Public Reason. *Philosophy & Technology* (2017). https://doi.org/10.1007/s13347-017-0263-5 arXiv:arXiv:1702.08608

[21] Reuben Binns. 2018. What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy, 2018* 16, 3 (2018), 73–80.

[22] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In *CHI '18.* https://doi.org/10.1145/3173574.3173951 arXiv:1801.10408

[23] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), 1–12. https://doi.org/10.1088/1742-5468/2008/10/P10008 arXiv:0803.0476

[24] Mark Bovens. 2007. Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal* 13, 4 (2007), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x arXiv:1468-0386

[25] Mark Bovens. 2010. Two concepts of accountability: Accountability as a virtue and as a Mechanism. *West European Politics* 33, 5 (2010), 946–967. https://doi.org/10.1080/01402382.2010.486119 arXiv:arXiv:1011.1669v3

[26] Mark Bovens, Thomas Schillemans, and Paul T. Hart. 2008. Does public accountability work? An assessment tool. *Public Administration* 86, 1 (2008), 225–242. https://doi.org/10.1111/j.1467-9299.2008.00716.x

[27] Mark Bovens and Stavros Zouridis. 2002. From to System Level -Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review* 62, 2 (2002), 174–184.

[28] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3 (2013), 209–227. https://doi.org/10.1007/s10676-013-9321-6

[29] Louis Brandeis. 1914. What publicity can do. In *Other people's money and how the bankers use it.* Frederick A. Srokes, New York. http://louisville.edu/law/library/special-collections/the-louis-d.-brandeis-collection/other-peoples-money-chapter-v

[30] Gijs Jan Brandsma and Thomas Schillemans. 2013. The accountability cube: Measuring accountability. *Journal of Public Administration Research and Theory* 23, 4 (2013), 953–975. https://doi.org/10.1093/jopart/mus034

[31] Kiel Brennan-Marquez. 2016. Plausible Cause: Explanatory Standards in the Age of Powerful Machines. *Ssrn* 1 (2016). https://doi.org/10.2139/ssrn.2827733

[32] Dennis Broeders, Erik Schrijvers, Bart van der Sloot, Rosamunde van Brakel, Josta de Hoog, and Ernst Hirsch Ballin. 2017. Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data. *Computer Law and Security Review* 33, 3 (2017), 309–323. https://doi.org/10.1016/j.clsr.2017.03.002

[33] Joanna Bryson and Alan Winfield. 2017. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer* 50, 5 (2017), 116–119. https://doi.org/10.1109/MC.2017.154

[34] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. *Proceedings of the ACM Conference on Intelligent User Interfaces* (2012), 169–178. https://doi.org/10.1145/2166966.2166996

[35] Jenna Burrell. 2015. How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms. *Ssrn* June (2015), 1–12. https://doi.org/10.2139/ssrn.2660674 arXiv:arXiv:1307.4531v1

[36] Robyn Caplan and danah Boyd. 2018. Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society* 5, 1 (2018), 205395171875725. https://doi.org/10.1177/2053951718757253

[37] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. 2018. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics* 24, 2 (2018), 505–528. https://doi.org/10.1007/s11948-017-9901-7

[38] Raja Chatila, Kay Firth-Butterflied, John C. Havens, and Konstantinos Karachalios. 2017. The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. *IEEE Robotics and Automation Magazine* 24, 1 (2017), 110. https://doi.org/10.1109/MRA.2017.2670225

[39] Kenneth Ward Church. 2017. Emerging trends: I did it, I did it, I did it, but. *Natural Language Engineering* 23, 3 (2017), 473–480. https://doi.org/10.1017/S1351324917000067

[40] Danielle Citron and Frank Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89, 1 (2014), 1–34. https://doi.org/10.3868/s050-004-015-0003-8 arXiv:arXiv:1011.1669v3

[41] Cary Coglianese and David Lehr. 2017. Regulating by robot: Administrative decision making in the Machine-learning era. *Georgetown Law Journal* 105, 5 (2017), 1147–1223. https://doi.org/10.1088/1748-0221/11/10/C10012

[42] Kate Crawford. 2016. Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science Technology and Human Values* 41, 1 (2016), 77–92. https://doi.org/10.1177/0162243915589635 arXiv:arXiv:1011.1669v3

[43] K. Crawford and T. Gillespie. 2014. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* (2014), 1461444814543163–. https://doi.org/10.1177/1461444814543163

[44] Andrej Dameski. 2018. A comprehensive ethical framework for AI entities: Foundations. In *11th International Conference Artificial General Intelligence 2018.* 42–51. https://doi.org/10.1007/978-3-319-97676-1_5

[45] John Danaher. 2016. The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology* 29, 3 (2016), 245–268. https://doi.org/10.1007/s13347-015-0211-1

[46] John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, Maria Helen Murphy, Niall O'Brolchain, Burkhard Schafer, and Kalpana Shankar. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4, 2 (2017), 205395171772655. https://doi.org/10.1177/2053951717726554

[47] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[48] Paul B. de Laat. 2017. Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology* (2017). https://doi.org/10.1007/s13347-017-0293-z

[49] S Dekker. 2018. Transparantie van algoritmes in gebruik bij de overheid.

[50] Lina Dencik, Arne Hintz, Joanna Redden, and Harry Warne. 2018. *Data Scores as Governance: Investigating uses of citizen scoring in public services.* Technical Report. Data Justice Lab, Cardiff. https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf

[51] Nicholas Diakopoulos. 2015. Accountability in Algorithmic Decision-making: A view from computational journalism. *ACM Queue* december (2015), 1–24.

[52] Nicholas Diakopoulos. 2015. Algorithmic Accountability. *Digital Journalism* 3, 3 (2015), 398–415. https://doi.org/10.1080/21670811.2014.976411

[53] Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3, 3 (2015), 398–415. https://doi.org/10.1080/21670811.2014.976411

[54] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision-making: A view from computational journalism. *Commun. ACM* 59, 2 (2016), 56–62. https://doi.org/10.1145/2844110

[55] Ezekiel Dixon-Román. 2016. Algo-Ritmo: More-Than-Human Performative Acts and the Racializing Assemblages of Algorithmic Architectures. *Cultural Studies <-> Critical Methodologies* 16, 5 (2016), 482–490. https://doi.org/10.1177/1532708616655769

[56] J.E. Dobson. 2015. Can An Algorithm Be Disturbed?: Machine Learning, Intrinsic Criticism, and the Digital Humanities. *College Literature: A Journal of Critical Literay Studies* 42, 4 (2015), 543–564. https://doi.org/10.1353/lit.2015.0037

[57] Danilo Doneda and Virgilio A.F. Almeida. 2016. What Is Algorithm Governance? *IEEE Internet Computing* 20, 4 (2016), 60–63. https://doi.org/10.1109/MIC.2016.79

[58] Cat Drew. 2016. Data science ethics in government. *Philosophical Transactions of the Royal Society* 374, 2083 (2016), 20160119. https://doi.org/10.1098/rsta.2016.0119

[59] Marina Drosou, H.V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in Big Data: A Review. *Big Data* 5, 2 (2017), 73–84. https://doi.org/10.1089/big.2016.0054

[60] ECP Platform voor de InformatieSamenleving. 2018. Artificial Intelligence Impact Assessment.

[61] Lilian Edwards and Michael Veale. 2017. Enslaving the algorithm : from a 'right to an explanation' to a 'right to better decisions'? (2017), 12 pages.

[62] Amitai Etzioni and Oren Etzioni. 2017. Incorporating Ethics into Artificial Intelligence. *Journal of Ethics* 21, 4 (2017), 403–418. https://doi.org/10.1007/s10892-017-9252-2

[63] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, New York.

[64] European Union. 2016. Regulation 2016/679 of the European parliament and the Council of the European Union. arXiv:arXiv:1011.1669v3

[65] Katherine Fink. 2018. Opening the government's black boxes: freedom of information and algorithmic accountability. *Information Communication and Society* 21, 10 (2018), 1453–1471. https://doi.org/10.1080/1369118X.2017.1330418

[66] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM SIGCHI Bulletin* 14, 3 (1996), 330 –347. https://doi.org/10.1145/249170.249184

[67] Urs Gasser and Virgilio A.F. Almeida. 2017. A Layered Model for AI Governance. *IEEE Internet Computing* 21, 6 (2017), 58–62. https://doi.org/10.1109/MIC.2017.4180835

[68] Gemeente Rotterdam. 2018. Betalingsregeling gemeentelijke belastingschuld. https://www.rotterdam.nl/loket/betalingsregeling-belastingschuld/

[69] Jen Jack Gieseking. 2018. Operating anew: Queering GIS with good enough software. *Canadian Geographer* 62, 1 (2018), 55–66. https://doi.org/10.1111/cag.12397

[70] A. Goffey. 2008. Algorithm. In *Software Studies: a lexicon*, Matthew Fuller (Ed.). MIT Press, Cambridge/London, 15–20.

[71] Mika Gröndahl, Keith Collins, and James Glanz. 2019. The dangerous flaws in Boeing's automated system. https://www.nytimes.com/interactive/2019/03/29/business/boeing-737-max-8-flaws.html

[72] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey Of Methods For Explaining Black Box Models. *Comput. Surveys* 51, 5 (2018). arXiv:1802.01933 http://arxiv.org/abs/1802.01933

[73] Andrew J. Hawkins. 2019. Deadly Boeing crashes raise questions about airplane automation. https://www.theverge.com/2019/3/15/18267365/boeing-737-max-8-crash-autopilot-automation

[74] Paul Henman. 2017. The computer says 'DEBT': Towards a critical sociology of algorithms and algorithmic governance. In *Data for Policy 2017: Government by Algorithm?* London. https://doi.org/10.5281/ZENODO.884117

[75] Paul Henman. 2019. Of algorithms, Apps and advice: digital social policy and service delivery. *Journal of Asian Public Policy* 12, 1 (2019), 71–89. https://doi.org/10.1080/17516234.2018.1495885

[76] Marc Hijink. 2018. Algoritme voorspelt wie fraude pleegt bij bijstandsuitkering. , 8–10 pages. https://www.nrc.nl/nieuws/2018/04/08/algoritme-voorspelt-wie-fraude-pleegt-bij-bijstandsuitkering-a1598669

[77] Russell Hotten. 2015. Volkswagen: The scandal explained. https://www.bbc.com/news/business-34324772

[78] Husson University. 2019. What is the software development life cycle? https://online.husson.edu/software-development-cycle/

[79] IEEE. [n. d.]. IEEE Code of Ethics. https://www.ieee.org/about/corporate/governance/p7-8.html

[80] Lucas D Introna. 2016. Algorithms , Governance , and Governmentality : On Governing Academic Writing. *Science Technology and Human Values* 41, 1 (2016), 17–49. https://doi.org/10.1177/0162243915587360

[81] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* 9, 6 (2014), 1–12. https://doi.org/10.1371/journal.pone.0098679 arXiv:arXiv:1209.0748v1

[82] Marijn Janssen and George Kuk. 2016. The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly* 33, 3 (2016), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

[83] Deborah G. Johnson and Helen Nissenbaum. 1995. *Computers, ethics & social values.* Prentice-Hall, Upper Saddle River.

[84] Natascha Just and Michael Latzer. 2017. Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, Culture and Society* 39, 2 (2017), 238–258. https://doi.org/10.1177/0163443716643157

[85] Christian Katzenbach. 2012. Technologies as Institutions: Rethinking the Role of Technology in Media Governance Constellations. In *Trends in Communication Policy Research, New Theories, Methods & Subjects*, Manuel Puppis and Natascha Just (Eds.). Intellect Books, Bristol, 117–138. https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[86] Jakko Kemper and Daan Kolkman. 2018. Transparent to whom? No algorithmic accountability without a critical audience. *Information Communication and Society* (2018), 1–16. https://doi.org/10.1080/1369118X.2018.1477967

[87] Rob Kitchin. 2019. The ethics of smart cities. https://www.rte.ie/brainstorm/2019/0425/1045602-the-ethics-of-smart-cities/

[88] Donald E. Knuth. 1984. Literate Programming. *Computers and Chemical Engineering* 22, 12 (1984), 1745–1747. https://doi.org/10.1016/S0098-1354(98)00029-5 arXiv:arXiv:1011.1669v3

[89] Utku Köse. 2017. Are We Safe Enough in the Future of Artificial Intelligence? A Discussion on Machine Ethics and Artificial Intelligence Safety. In *Scientific Methods in Academic Research and Teaching International Conference*. 184–197.

[90] Felicitas Kraemer, Kees Van Overveld, and Martin Peterson. 2011. Is there an ethics of algorithms ? *Ethics and Information Technology* 13, 3 (2011), 251–260. https://doi.org/10.1007/s10676-010-9233-7

[91] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165 (2017), 633–705. https://doi.org/10.3868/s050-004-015-0003-8 arXiv:arXiv:1011.1669v3

[92] Leer- en Expertisepunt Open Overheid. 2019. Actieplan Open Overheid 2018-2020. https://www.open-overheid.nl/actieplan-open-overheid-2018-2020-open-moet-het-zijn/

[93] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2017. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* (2017). https://doi.org/10.1007/s13347-017-0279-x

[94] Lawrence Lessig. 1999. *Code: and other laws of cyberspace.* Basic Books, New York.

[95] Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P.A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine* 6, 7 (2009). https://doi.org/10.1371/journal.pmed.1000100 arXiv:arXiv:1011.1669v3

[96] Michael Lipsky. 1980. *Street-level bureaucracy: dilemmas of the individual in public services.* Russell Sage Foundation, New York.

[97] Katharina Loeber. 2018. Big Data, Algorithmic Regulation, and the History of the Cybersyn Project in Chile, 1971–1973. *Social Sciences* 7, 4 (2018), 65. https://doi.org/10.3390/socsci7040065

[98] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic Authority: The Ethics, Politics, and Economics of Algorithms that Interpret, Decide, and Manage. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16* (2016), 1057–1062. https://doi.org/10.1145/2851581.2886426

[99] Lei Ma, Zhongqiu Zhang, and Nana Zhang. 2018. Ethical Dilemma of Artificial Intelligence and its Research Progress. In *IOP Conference Series: Materials Science and Engineering*, Vol. 392. https://doi.org/10.1088/1757-899X/392/6/062188

[100] Vidushi Marda. 2018. Artificial intelligence policy in India: A framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2018). https://doi.org/10.1098/rsta.2018.0087

[101] Kirsten Martin. 2018. Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 0, 0 (2018), 1–16. https://doi.org/10.1007/s10551-018-3921-3

[102] James McGrath and Ankur Gupta. 2018. Writing a Moral Code: Algorithms for Ethical Reasoning by Humans and Machines. *Religions* 9, 8 (2018), 240–259.

[103] Dan McQuillan. 2018. People's Councils for Ethical Machine Learning. *Social Media and Society* 4, 2 (2018), 1–10. https://doi.org/10.1177/2056305118768303

[104] Eden Medina. 2015. Rethinking algorithmic regulation. *Kybernetes* 44, 6-7 (2015), 1005–1019. https://doi.org/10.1108/K-02-2015-0052

[105] Thomas Metzinger. 2019. Ethics washing made in Europe. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

[106] 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), 205395171667967. https://doi.org/10.1177/2053951716679679

[107] Annemarie Mol. 2002. *The body multiple: ontology in medical practice.* Duke University Press, Durham.

[108] Joshua New and Daniel Castro. 2018. *How Policymakers can foster Algorithmic Accountability.* Technical Report. Center for Data Innovation, Washington.

[109] New York City. 2018. Automated Decision Systems Task Force. https://www1.nyc.gov/site/adstaskforce/index.page

[110] Daniel Neyland. 2016. Bearing Account-able Witness to the Ethical Algorithmic System. *Science, Technology, & Human Values* 41, 1 (2016), 50–76. https://doi.org/10.1177/0162243915598056

[111] Daniel Neyland and Norma Möllers. 2017. Algorithmic IF ... THEN rules and the conditions and consequences of power. *Information Communication and Society* 20, 1 (2017), 45–62. https://doi.org/10.1080/1369118X.2016.1156141

[112] Helen Nissenbaum. 1994. Computing and accountability. *Commun. ACM* 37, 1 (1994), 72–80. https://doi.org/10.1145/175222.175228

[113] Dietmar Offenhuber. 2017. *Waste is information: infrastructure legibility and governance.* MIT Press, Cambridge/London.

[114] Partnership on AI. 2019. About. https://www.partnershiponai.org/about/

[115] Partnership on AI. 2019. Partnership on AI. https://www.partnershiponai.org/

[116] Frank Pasquale. 2015. Digital Star Chamber. https://aeon.co/essays/judge-jury-and-executioner-the-unaccountable-algorithm

[117] David J. Pauleen, David Rooney, and Ali Intezari. 2017. Big data, little wisdom: trouble brewing? Ethical implications for the information systems discipline. *Social Epistemology* 31, 4 (2017), 400–416. https://doi.org/10.1080/02691728.2016.1249436

[118] Wolter Pieters. 2011. Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology* 13, 1 (2011), 53–64. https://doi.org/10.1007/s10676-010-9253-3

[119] Craig Plain. 2007. Build an Affinity for K-J Method. *Quality Progress* 40, 3 (2007), 88.

[120] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14. https://doi.org/10.1007/s10676-017-9430-8 arXiv:1707.07232

[121] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, 1135–1144. arXiv:1602.04938 http://arxiv.org/abs/1602.04938

[122] Rijksoverheid. 2014. Besluit SUWI. https://wetten.overheid.nl/BWBR0013267/2019-01-01{#}Hoofdstuk5a

[123] Alex Rosenblat, Tamara Kneese, and Danah Boyd. 2014. Algorithmic Accountability. In *The Social, Cultural & Ethical Dimensions of "Big Data"*. Data & Society Research Institute, New York. https://datasociety.net/pubs/2014-0317/AlgorithmicAccountabilityPrimer.pdf

[124] Florian Saurwein, Natascha Just, and Michael Latzer. 2015. Governance of algorithms: Options and limitations. *Info* 17, 6 (2015), 35–49. https://doi.org/10.1108/info-05-2015-0025

[125] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (2017), 205395171773810. https://doi.org/10.1177/2053951717738104

[126] Security.nl. 2018. Kamervragen over algoritmen voor opsporen bijstandsfraude. https://www.security.nl/posting/557836/Kamervragen+over+algoritmen+voor+opsporen+bijstandsfraude?channel=rss

[127] Bernd Carsten Stahl and David Wright. 2018. Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security and Privacy* 16, 3 (2018), 26–33. https://doi.org/10.1109/MSP.2018.2701164

[128] Stats NZ. 2018. *Algorithm Assessment Report.* Technical Report. Stats NZ, Wellington.

[129] Daniel Susser. 2019. Ethics Alone Can't Fix Big Tech. https://slate.com/technology/2019/04/ethics-board-google-ai.html

[130] Jim Torresen. 2018. A Review of Future and Ethical Perspectives of Robotics and AI. *Frontiers in Robotics and AI* 4, January (2018). https://doi.org/10.3389/frobt.2017.00075

[131] Totta Data Lab. 2019. Gemeentelijke Fraudedetectie. https://www.tottadatalab.nl/portfolio-item/gemeentelijke-fraudedetectie/

[132] T. Van Ark. 2018. Kamervraag/vragen van het lid Buitenweg (GroenLinks).

[133] Iris Van der Tuin and Rick Dolphijn. 2012. "Matter feels, converses, suffers, desires, yearns and remembers" Interview with Karen Barad. In *New Materialism: Interviews & Cartographies.* Open Humanities Press, Ann Arbor, 48–70.

[134] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. (2018), 1–14. https://doi.org/10.1145/3173574.3174014 arXiv:1802.01029

[135] Anton Vedder and Laurens Naudts. 2017. Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers and Technology* 31, 2 (2017), 206–224. https://doi.org/10.1080/13600869.2017.1298547

[136] Ben Wagner. 2016. Algorithmic regulation and the global default: Shifting norms in Internet technology. *Etikk i Praksis* 10, 1 (2016), 5–13. https://doi.org/10.5324/eip.v10i1.1961

[137] Ben Wagner. 2018. Ethics as an Escape from Regulation: From 'ethics-washing' to ethics-shopping? In *Being Profiled, Cogitas Ergo Sum*, E. Bayamliogu, I. Baraliuc, L.A.W. Janssens, and M. Hildebrandt (Eds.). Amsterdam University Press, Amsterdam, 84–90.

[138] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018.* Technical Report. AI Now/New York University, New York.

[139] Ben Williamson. 2015. Governing software: networks, databases and algorithmic power in the digital governance of public education. *Learning, Media and Technology* 40, 1 (2015), 83–105. https://doi.org/10.1080/17439884.2014.924527

[140] Karen Yeung. 2017. Algorithmic regulation: A critical interrogation. *Regulation & Governance* April (2017), 1–19. https://doi.org/10.1111/rego.12158

[141] Karen Yeung. 2017. 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication and Society* 20, 1 (2017), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

[142] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. In *IJCAI International Joint Conference on Artificial Intelligence*, Vol. 2018-July. 5527–5533. https://doi.org/10.24963/ijcai.2018/779 arXiv:1812.02953v1

[143] 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 and Human Values* 41, 1 (2016), 118–132. https://doi.org/10.1177/0162243915605575

[144] Malte Ziewitz. 2016. Governing Algorithms: Myth, Mess, and Methods. *Science Technology and Human Values* 41, 1 (2016), 3–16. https://doi.org/10.1177/0162243915608948

A VIGNETTES

Below several vignettes are presented, which provide some more concrete illustration of the theory and problems described above.

A.1 Checking repayment arrangements

The municipality of Rotterdam [68] has a simple rule-based system which checks whether or not people live up to their repayment arrangement. Deviating one cent from the agreed upon installment automatically terminates the arrangement. While this is certainly a legal way to implement such as system, it may not be the most compassionate. Interestingly, the municipality of Rotterdam transitioned with the municipal election of 2018 from a center-right coalition (Leefbaar Rotterdam, CDA, and D66) to a coalition which also encompasses left-leaning parties (GroenLinks, VVD, D66, PvdA, CDA, and CU-SGP).

A.2 Automatic anonymization

One of the Netherlands' four largest municipalities is currently training a system, build by a third party, to automatically anonymize permits, so they can, eventually, make these available to the public pro-actively. The system is part of an effort to minimize spending on personnel which, currently, manually needs to remove personally identifiable information from the documents. At the moment, there is still a human agent which monitors and corrects the system's output, but the municipality is deliberating to automate the system (i.e. remove the human from the loop) when the system reaches 95% accuracy.[4] This level of accuracy means that in 2.5% of all cases the system may have removed too much information from the document, and in 2.5% of instances personally identifiable information may not have been removed thoroughly.

A.3 Fraud detection

A municipality in the east of the Netherlands, with approximately 100.000 inhabitants, works together with a third party in a pilot in which they try to detect fraud amongst people who receive social benefits.[5] The municipality is quite conscious about the dangers of such algorithmic assessments. As such, they deliberately place themselves not in the front lines of data-driven developments, but rather want to learn from others' best practices.

The municipality's biggest concern with this system is that they do not know how particular aspects of the system work (e.g. what particular kind of weighting is used for what parameters), and thus cannot take full accountability for the system. They are very conscious about this problem and thus designed an exploratory space (i.e. the pilot) in which they deliberately chose to not let the system's results be a new informational category for the investigatory process, but to treat it as any other anonymous tip. However, the team noted their dissatisfaction with the current setup, and are planning to request access and/or insight in the algorithm. The investigator noted that they know that the system works, but they have no idea why it works. They also want to consult other municipalities who work with the same firm, to discuss how they tackled this problem.

[4]Fieldnotes: November 11th 2018.
[5]Fieldnotes: December 12th 2018.

A.4 SyRI (System Risk Indicator)

Several municipalities (and other public sector organizations) in the Netherlands have used the System Risk Indicator (SyRI). SyRI is used to assess which people on social benefits are more likely to commit fraud, and considers a vast amount of data: from one's water usage to which permits they have requested [18, 122]. The system sparked a lot of upheaval in the Netherlands as municipalities started using it as a way to pro-actively screen their citizens [76, 126, e.g.].

The system officially reports persons suspecting fraud to the minister in charge of the Ministry of Social Affairs and Employment. The minister delegates this task in turn to civil servants of, for instance, the respective municipalities using the system in their investigations. Nevertheless, it is the minister and/or their undersecretary who are/is held accountable in the political forum of the Dutch House of Representatives [132, e.g.].

Much details about the system are undisclosed. Because of this, a group of civil society organisations, united under the name 'Bij Voorbaat Verdacht' (tr. 'Suspect by default'), tries to uncover how the system works. They have submitted FOI requests which were partially successful, and are, at the time of writing, suing the government for openness about the system. Their argument is that SyRI has no place in a democratic environment where civilians are required to share their data with governmental parties, which are subsequently connected and used for preventive profiling measures of which the civilian is uninformed and which they cannot question due to the system's opacity [19]. The government, in their turn, argue that exposing the modus operandi of the system may lead to gaming effects, thus making such sensitive aspects of the system transparent would be ill-advised [132].

B METHODOLOGY

B.1 Query design

In order to accommodate the diversity of studies relating to the topic of algorithmic accountability, relevant associated terms need to be identified. This is done using a recursive query design (see figure 1.). The recursivity lies in the repetition of steps and their subsequent snowball effect. First, an exploratory query is designed, based on the relational strength of the keywords of 27 pre-identified articles. Using this exploratory query we then collect new articles, and from those relevant we again extract keywords and assess their strength, creating a preliminary query. This preliminary query leads up to the creation of the final query.

Some justification is needed with regard to the procedure. Author-identified keywords added to academic articles were chosen as an indicator of relatedness. Keywords were chosen as indicator as they are created to briefly represent the content of an article and to be specific and legible to one's field of study. In mapping the relations between keywords, oft-discussed themes should come to the fore, as well as the diversity of perspectives and terms between disciplines.

After the keywords were inventoried, they were made more generalizable, by using the * operator. After this, colocation of keywords were identified and mapped using network visualisation tool Gephi [14]. The keywords which related the strongest to one another informed the new query.

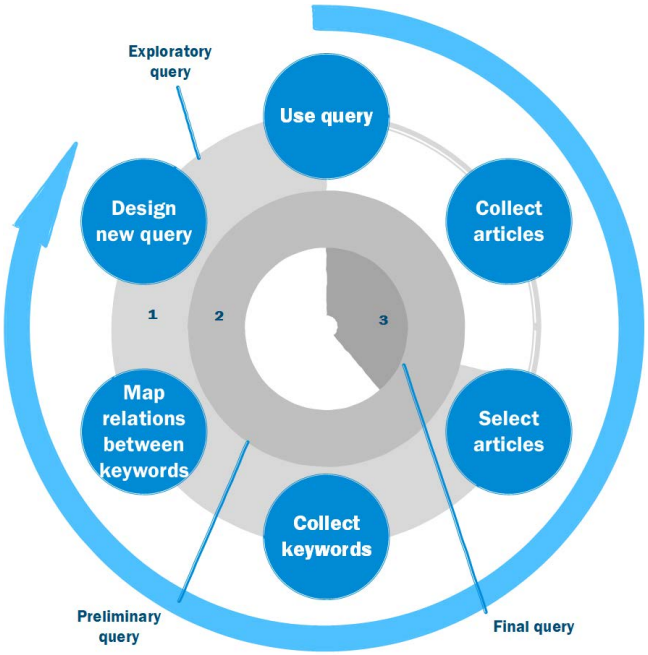

Figure 1: Representation of the recursive query design process.

Using the resulting definitive query, the corpus was selected. This material was selected by screening titles and abstracts for their relevance to the topic, and their adherence to the eligibility criteria. As computational systems tend to become obsolete quite quickly, this study will cover the last ten years (2008 up to and including 2018). With an eye on future replicability of the study, the review will limit itself to those publications published in English. Only works that have been published will be reviewed (e.g. working papers will not be included). Only articles that present original academic work will be included (e.g. research article, review article), whereas, for instance, introductions were excluded.

B.1.1 Exploratory query design. 27 academic articles [8, 9, 17, 20, 22, 28, 35, 48, 52, 58, 61, 65, 86, 90, 93, 98, 106, 110, 111, 120, 124, 135, 139–141, 143, 144] which were found to be relevant to the topic prior to the start of the systematic review were assessed for their keywords. These articles were all strongly connected to these of algorithmic accountability, explainability/transparency, ethics, decision-making, and governance. Books, reports and academic articles without keywords were excluded from this exploratory inventory. As some keywords overlapped partially (e.g. algorithmic decision-making/algorithmic decision making/automated decisions/automated decisions) keywords were grouped together when overlap occurred (e.g. 'decision*'). In total, 79 keywords were found after resolving this overlap. Next, for each keyword, of the article relation between the other keywords of the article was mapped. This lead to an inventory of 879 relations, or 'edges'.

These edges were subsequently fed into the network visualization program Gephi [14]. Among the 879 edges, 752 unique ones were found, meaning there are 127 instances in which different articles use the same keywords. The colocations were visualized using the ForceAtlas2 algorithm [81]. From this exploration, those colocated keywords were selected which had the highest degree (both in- and out-degree). The threshold was set at degree >= 35, meaning that in order to be considered for the next step in building the query design, these colocated keywords had to have a sum of incoming/outgoing connections equal to or greater than 35 (fig. 2.). This left 9 nodes (11.39%), and 57 edges (7.58%).

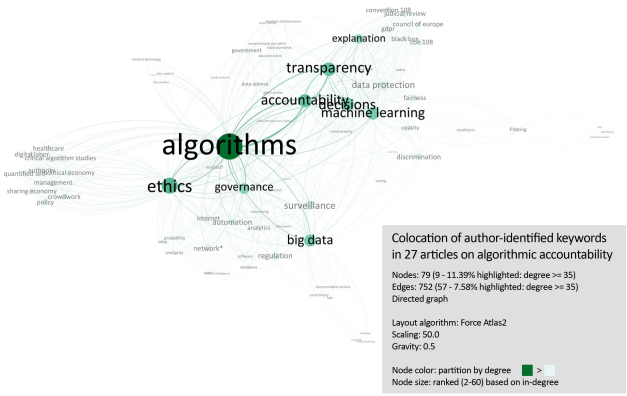

Figure 2: Exploratory mapping of colocations of keywords in articles on algorithmic accountability, filtered on degree >= 35.

This selection was subsequently used to build the query. The edges table of the filtered subset was exported, and duplicate relations were added together. The result was ordered on edge weight (i.e. how strong/frequent the colocation is). These insights, together with the mapping of the colocations, allow for a first, considered query design. Whereas the combined edge weight conveys the strength/frequency of the relation between the terms. The network graph gives on indication of the discourses which draw on particular keywords.

After generating insight in the strength of the relations between the keywords, the strongest colocated keywords were selected for the query design. To this end, edge weights <= 4 were not included. However, where further specification of the query was preferable, as some terms could be quite general (e.g. 'big data'), they were used to supplement the query. The query that was designed based in the selection is as follows:

> ["algorithmic accountability" OR algorithm* AND accountabl* OR algorithm* AND accountabl* AND transparency OR governance AND algorithm* OR algorithm* AND transparency OR ethic* AND algorithm* OR transparency AND decision* AND algorithm* OR algorithm* AND "big data" AND governance OR algorithm* AND "big data" AND decision* OR transparency AND "machine learning" AND algorithm* OR transparency AND explanation AND algorithm* OR accountabl* AND decision* AND algorithm*]

B.1.2 Designing the final query. This initial query was used to gather more relevant publications, and to subsequently finetune the query design. The exploratory query was used to search for more relevant articles, and those articles were then again used for a colocation mapping strategy in order to improve the query design. In other words, the query design is circular, so that potential bias emanating from the first batch of 27 articles which were seen as relevant, could be nullified – and initial assumptions about important keywords could be tested.

As articles were best suited to this approach – as they often include author-specified keywords – both Web of Science and SCOPUS were queried. Querying titles, keywords and abstracts in SCOPUS delivered 7.019 results in total. The search was then further specified for the period 2008-2018 (5.397 results), to include only English papers (5.145 results). Web of Science allows for searching a 'topic', which – similar to SCOPUS – encompasses the title, keywords, and abstract of a given work. Querying topics in the Web of Science resulted in 2.127 results for the period 2008-2018, of which 2.076 were English. As these results needed to be screened manually, the smaller corpus of Web Of Science was used for this second exploration, and the search results were exported as TSV files.

After their export, the files were cleaned. All data entries were screened using the same procedure. First, the title was checked for its relevance. An article is considered relevant if algorithmic accountability is the main topic of the publication. If the title was found to be relevant, the article was included, if not, it was excluded. In case of doubt, the abstract was assessed for its relevance, following the same procedure. If doubt still remained after reading through the abstract, the publication was included provisionary (see fig. 4 for an overview of the entire process).

This screening resulted in 114 inclusions (5.5%), 99 provisional inclusions (4.8%), and 1.863 exclusions (89.7%). Thus, 10.3% of the results were found to be of (potential) interest. The 114 inclusions provided the basis for a second round of colocation mapping. Of the 114 papers, 25 (22%) articles were found to have no keywords, leaving 89 (78%) articles which did include such keywords. in which 270 keywords and 2,870 colocations were identified. Again, these relations were investigated using Gephi.

The network appeared to be connected, except for one paper [69], whose keywords did not overlap with any of the other articles. This single paper was excluded from consideration for the subsequent query design process. It was filtered out by using a Giant Component filter (98.15% of nodes and 99.3% of edges visible). The remainder of the connections were mapped using ForceAtlas2 [81]. Modularity [23] was exploratively used with different resolutions (displayed in fig. 3: resolution 1.0) to see if keyword preferences amongst the various disciplines could be detected, but did not produce such results.

Subsequently, the edges table was exported and the weights were combined as described earlier. As this second round of enveloped a greater number of relations – the cutoff point was not set at 4, but rather at 10. Thus, relations with a combined edge weight >= 11 were included in the final query design.

Using the combined edge weight, a new query was designed. As before, excluded terms might be used to complement very general terms where necessary. The final constructed query is as follows:

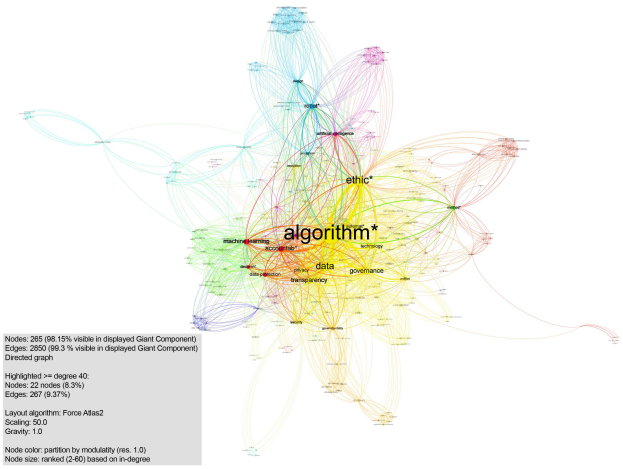

Figure 3: Exploratory mapping of colocations of keywords in articles on algorithmic accountability, filtered on degree >= 40.

> ["algorithmic accountability" OR algorithm* AND ethic* OR algorithm* AND data AND ethic* OR algorithm* AND data AND transparency OR algorithm* AND data AND accountab* OR algorithm* AND governance OR algorithm* and accountab* OR algorithm* AND transparency OR algorithm* AND technology AND transparency OR algorithm* AND technology AND ethic* OR algorithm* AND technology AND accountab* OR algorithm* AND privacy AND transparency OR transparency AND accountab* AND algorithm* OR ethic* AND "artificial intelligence" OR algorithm* AND automat* AND decision* OR algorithm* AND "machine learning" AND transparency OR algorithm* AND machine learning" AND ethic*]

B.2 Information sources

Using the specified query, SCOPUS, and Web of Science were searched on November 8th 2018. Similarly to the procedure in the query design stage, the databases were queried for the period 2008-2018, and only publications in English were included. Querying Web of Science generated 5,731 results for the period 2008-2018, of which 5,618 were in English. Due to SCOPUS' limitation on the downloading of the complete information (a maximum of 2,000 entries at a time), the database had to be queried for each additional query separately and sometimes even had to be split per year. The separate files were taken together afterwards. Querying SCOPUS resulted in 19,892 hits for the period 2008-2018, of which 19,033 were in English. As the query was broken down, 2,845 duplicates had to removed, leaving 16,188 unique titles.

The 5,618 titles from Web of Science and the 16,188 titles from SCOPUS were subsequently manually assessed for their relevance following a similar procedure as per the query design stage (i.e. assessing relevance of the title/title and abstract). After this initial

round, the (provisionally) included titles were taken together. This resulted 264 (provisional) inclusions from SCOPUS, and 204 (provisional) inclusions from Web of Science. After merging both corpora, this resulted in 371 titles. Subsequently final decisions were made with regards to the provisionally included articles (34 excluded), and corpus was limited to journal and proceeding articles (e.g. no book reviews, introductions to special issues). This resulted in a final selection of 242 articles. The articles' sources were checked against Beall's list of predatory journals [16], but no predatory outlets were found among the selection. To prioritize and group the reading material a rudimentary affinity mapping [119] was done based on the titles and abstracts. In the present contribution, the 93 articles which were identified as 'core articles' (those articles that seemed to related the strongest to the topic) were analyzed and presented.

Of the 93 selected articles, 30 were excluded. Of these, 5 articles were not accessible to the author, even after requesting them from the respective authors. Seven were excluded because their focus found was not to be on algorithmic accountability upon reading the entire piece. 15 were excluded because they were found not to be original research articles (e.g. opinion pieces, commentary, introductions to special issues). Three were excluded as they did not contain results. Two were excluded for other reasons. This left 61 articles which were thematically analyzed. (see fig. 4.).

B.3 Limitations and further research

As the methodology adopted for this paper is innovative, there are some limitations and aspects to it that need further study. First of all, the methodological merits need to be evaluated and assessed in its own right. Second, the methodological approach needs to be scrutinized for its potential skewedness or bias. It may be possible that the approach, though designed to be as inclusive as possible, may disfavor particular communities implicitly (e.g. Global North being 'dominant' mode of conversing about this phenomenon, thus the recursive query design might disfavor work from the Global South which operates in a different discourse), or the initial batch of articles used to distill the exploratory query may have been skewed. While, we do find several papers with Global South perspectives in the corpus (e.g. [100, 104], this is something that we take in account when evaluating the methodology elsewhere.

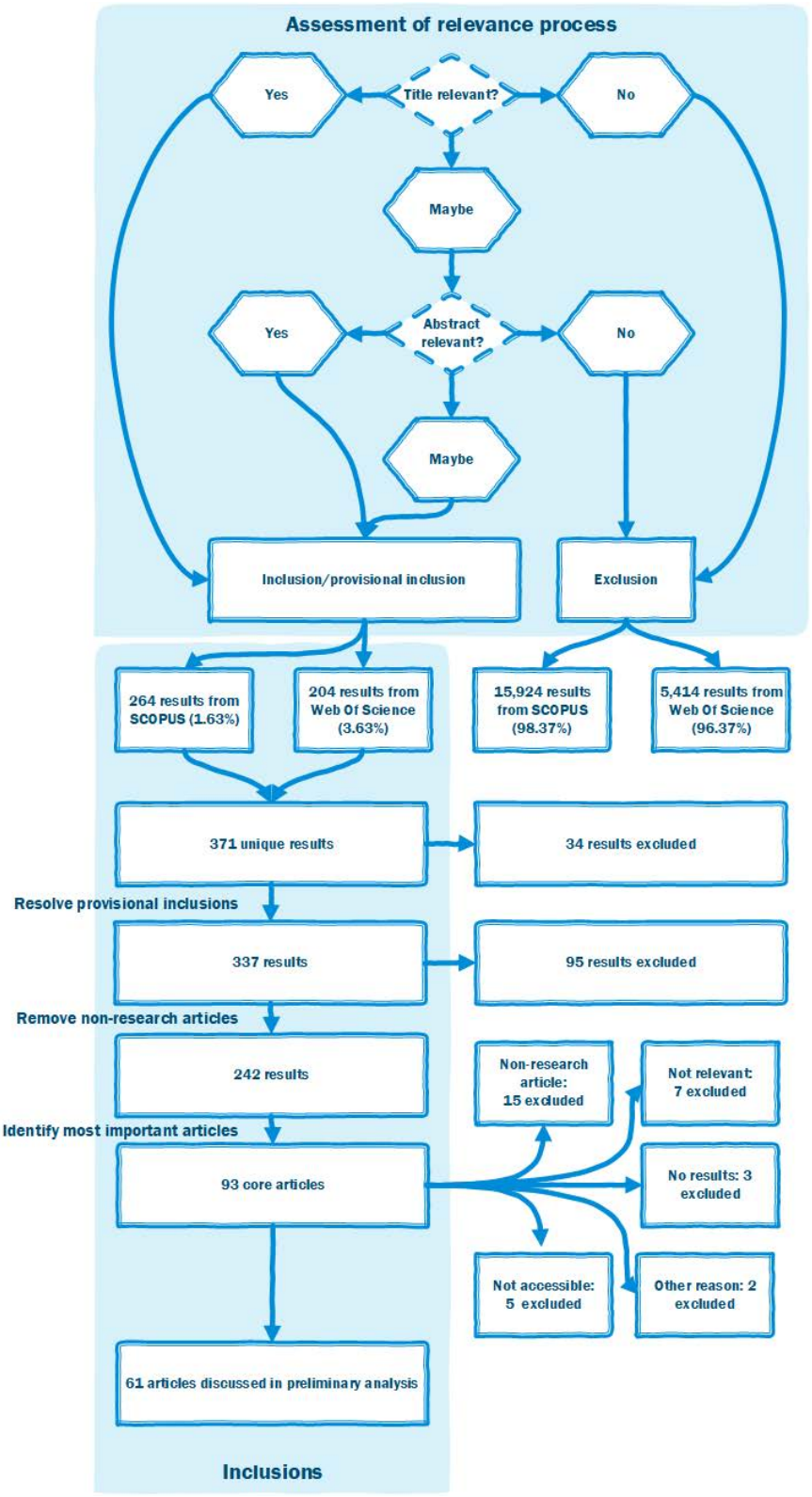

Figure 4: Flowchart of selection process.