Transparency International Anti-Corruption Helpdesk Answer

Algorithmic transparency and accountability

Authors: Niklas Kossow, Svea Windwehr and Matthew Jenkins

Reviewers: Daniel Eriksson and Jon Vrushi, Transparency International, and Laurence Millar, Transparency International New Zealand

Date: 05 February 2021

Computer algorithms are being deployed in ever more areas of our economic, political and social lives. The decisions these algorithms make have profound effects in sectors such as healthcare, education, employment, and banking. Their application in the anti-corruption field is also becoming increasingly evident, notably in the domain of anti-money laundering.

The expansion of algorithms into public decision making processes calls for a concomitant focus on the potential challenges and pitfalls associated with the development and use of algorithms, notably the concerns around potential bias. This issue is made all the more urgent by the accumulating evidence that algorithmic systems can produce outputs that are flawed or discriminatory in nature. The two main sources of bias that can distort the accuracy of algorithms are the developers themselves and the input data with which the algorithms are provided.

Equally troublingly, the analytical processes that algorithms rely on to produce their outputs are often too complex and opaque for humans to comprehend, which can make it extremely difficult to detect erroneous outputs. The Association for Computing Machinery (2017) points to three potential causes of opacity in algorithmic decision making processes. First, there are technical factors that can mean that the algorithm's outcomes may not lend themselves to human explanation, a problem particularly acute in machine-learning systems that can resemble a "black box." Second, economic factors such as commercial secrets and other costs associated with disclosing information can inhibit algorithmic transparency. Finally, socio-political challenges, such as data privacy legislation may complicate efforts to disclose information, particularly with regards to the training data used.

This paper considers these challenges to algorithmic transparency, and seeks to shed some light on what could constitute meaningful transparency in these circumstances, as well as how this can be used to leverage effective accountability in realm of algorithmic decision-making.

Given the potential for automated decision-making to result in discriminatory outcomes, the use of algorithms in public administration needs to come with certain standards. What these safeguards look like will vary in different contexts, but should be built into each stage of adopting algorithmic systems: from design, through the building and testing phases through to implementation (see Center for Democracy and Technology 2017).

Ultimately, institutions that use algorithms as part of automated decision making processes need to be held to the same standards as institutions in which humans make these decisions. Developers must ensure that the algorithmic systems they design are able to comply with the demands of impartial and accountable administration, such as accountability, redress and auditability.

This document should not be considered as representative of the Commission or Transparency International's official position. Neither the European Commission,Transparency International nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information.

The Anti-Corruption Helpdesk is operated by Transparency International and funded by the European Union.

Caveat

This paper considers the knotty topic of transparency and accountability in the use of algorithms. While it touches on the use of algorithms in the private sector to provide illustrative examples, it is primarily considered with the implications for public administration posed by the adoption of algorithmic decision-making. While noting that the use of algorithms is most prominent in the area of service delivery, the paper also considers the ramifications for the adoption of algorithms at higher levels of governmental policy-making.

The paper briefly reflects on ways in which algorithms can (re)produce opportunities for the abuse of entrusted power, including corruption, but does not analyse in detail the topic of potentially "corrupt" algorithms – that is algorithms developed with the specific aim of achieving fraudulent outcome. Neither does this paper seek to address the potential application of artificial intelligence and machine learning as anti-corruption tools, a matter that has been studied elsewhere (Adam and Fazekas 2018).

Contents

Main points

The OECD (2019a) identifies a number of opportunities and challenges in relation to the use of artificial intelligence (AI) and algorithms in the realm of good governance and anti-corruption work.

"Opportunities

— Algorithmic decision-making can identify and predict potential corruption issues by digesting diverse data sets.

— AI can increase the efficiency and accuracy of due diligence, and identify loopholes within regulatory frameworks.

Challenges

— The predictions and performance of algorithms are constrained by the decisions and values of those who design them, the training data they use, and their intended goals for use.

— By learning based on the data fed to them, AI-powered decisions face the risk of being biased, inaccurate or unfair especially in critical areas such as citizen risk profiling in criminal justice procedure and access to credit and insurance. This may amplify social biases and cause discrimination.

— The risk of inequality is exacerbated, with wealth and power becoming concentrated into a few AI companies, raising questions about the potential impacts of AI on income distribution and on public trust in government.

— The difficulty and sometimes technical impossibility of understanding how AI has reached a given decision inhibits the transparency, explainability and interpretability of these systems."

Algorithmic systems in public administration

Algorithmic systems that assist and often replace humans in decision-making processes are no longer a matter of science fiction; they have long since become a reality in many areas of our lives.

To many citizens, their first conscious encounter with such systems may seem novel, gimmicky or simply irksome. One such example is the expanding use of chatbots in services provided to citizens and customers. While the use of chatbots such as Bobbi, employed by the city of Berlin to help answer citizens' questions regarding the coronavirus, can improve service efficiencies (Puntschuh and Fetic 2020), almost half of people surveyed find these conversational language interfaces "annoying" (Artificial Solutions 2020).

Yet behind the scenes, the intensifying impact of digitalisation on our societies is also increasingly shaping the way we govern. As growing numbers of policymakers and administrators come to rely on artificial intelligence (AI) and algorithmic systems, we are in the middle of a profound change in how public administrations work and how public goods and services are administered. As digital instruments and repositories of governance applications proliferate (apolitical 2020), so too does the reliance on big data and ever-more sophisticated forms of data analysis. The result is a growing tendency to rely on data based decision making to guide policy (Höchtl, Parycek & Schöllhammer 2016; Athey 2017).

While advocates contend that the use of algorithms in public life can lead to more consistent outcomes by replacing fallible and irrational humans (Coglianese 2016*)*, there is mounting evidence that algorithmic systems can reproduce human foibles (Zerilli et al 2019).

Worse still, the premature rollout of flawed algorithmic decision-making processes can embed such biases at the heart of supposedly neutral systems, reinforcing, for instance, historic patterns of discrimination in areas such as law enforcement (MIT Technology Review 2020).

Furthermore, if purposefully designed to favour specific outcomes, algorithmic systems could also assist corrupt causes or actors, while giving the impression of neutral decision-making.

Against this backdrop, it is paramount to ensure that these tools are conducive to the common good, and that their use does not give rise to abuses of power that can result in corruption.

The rest of the paper proceeds as follows. The first section seeks to define algorithmic decision making systems and consider how opacity in their development and use could lead to corrupt practices. As highlighted in the second section, while greater transparency could (partially) address this dilemma, there are inherent characteristics of algorithmic systems that mean that transparency alone is no panacea to the forms of corruption they foster. Instead, the third section of the paper proposes a greater policy focus on accountability and transparency relating to the design, training, procurement and deployment of algorithmic systems. Ultimately, the paper contends that transparency alone is insufficient; we need to adopt a more holistic view of accountability in the context of algorithms.

What are algorithmic systems and how do they work?

Historically, 'government by algorithm' refers to the idea that with perfect access to high-quality data, governments could use complex algorithmic systems to make ideal decisions that factor in all relevant data points. This notion was especially popular with socialist governments that strove to automate planned economies (Morozov 2014b). One example is Project Cybersyn, through which Chilean president Salvador Allende attempted to construct a distributed decision making system to control and manage the Chilean economy.

Ultimately, as a result of the 1973 coup d'etat that marked the end of Allende's government, Cybersyn was abandoned before it could be fully put to use. Despite this, algorithmic systems have long since found their way into policy making and governance (Loeber 2018). Today, algorithms have a role in decision making processes that decide, for example, where fire stations are built, whether a person receives bail, or how public benefits eligibility is determined (AI Now Institute 2018).

Data collection and analysis has been a central part of governance for decades, albeit in different forms that might appear unrelated to today's data processing techniques. Yet, the current information age has brought with it a pressing need to handle and interpret ever-vaster amounts of data. As more and more data is produced at an exponential rate by individuals, industry and governments, many researchers, technologists and public administrators have come to see algorithmic systems as a necessary tool to cope with the sheer volume of data available. Increasingly, countries and businesses are pointing to the potential economic and social benefits to be reaped through the more meaningful use of the mass of data available.

In particular, the advocates of applying increasingly sophisticated data science and data analysis techniques to political life underscore the advantages that improved understanding, control and security of societal and economic forces could bring (Mohabbat Kar, Thapa, Parycek 2018). At the same time, there is growing societal unease about the increased centrality and authority accorded to algorithmic systems, as well as the lack of transparency in their application (Kroll et al 2017).

What are algorithmic systems?

An algorithmic system is a set of algorithms. According to the Association for Computing Machinery (2017), an algorithms is "a self-contained step-by-step set of operations that computers and other 'smart' devices carry out to perform calculation, data processing, and automated reasoning tasks. Increasingly, algorithms implement institutional decision-making based on analytics, which involves the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming, and operations research to quantify performance."

At heart, algorithms define a set of steps that must be followed in order to achieve a specified result. The algorithms that power almost every piece of software dictate the steps a computer follows to make a decision or execute a task (Domingos 2015). Algorithms are thus found in countless applications used in daily life, ranging from such mundane tools as calculators, to more complex ones like systems laptops, online maps or search engines. Most applications comprise several algorithms pursuing different tasks and functions, making up algorithmic systems.

Beyond this purely mathematical or computational term, algorithms have come to be perceived as a broader category of technology that can take (usually reliable) decisions based on more or less complex rules and using available data (Mittelstadt et al 2016). As such, algorithms, or algorithmic decision systems are components of technologies that are considered to be, or contain, artificial intelligence (AI). AI refers to the capability of a computer to imitate intelligent human behaviour, and also describes the field of computer science dedicated to simulation of intelligent behaviour in machines (Merriam-Webster 2020). It is worth noting, however, that what is considered to be AI is a shifting frontier: few people would think of their computer's calculator application as AI, although it completes complex tasks based on predefined rules more reliably than humans.

Very broadly, two categories of algorithmic systems can be distinguished: rule-based and learning systems. Rule-based algorithms follow a simple "if → then" logic, which limits their application as every instruction must be predefined. Such systems are usually not capable of dealing with new information or unforeseen problems.

Learning systems, also called machine learning systems, are systems that improve through experience. For example, such a system could be fed with training data consisting of pictures labelled as "cat" and "no cat". Over time, and without being told to do so, that system would learn how to identify cats and distinguish from other animals. Machine learning systems can either be supervised - the computer is presented with example inputs and desired outputs - or unsupervised - the computer is not given any labels and is left to discover patterns in the data it is given on its own.

Crucially, the design process of algorithmic decision systems - understood as the definition of the problem to be solved, the development of a model to solve the problem and the subsequent selection or creation of the appropriate algorithm to execute the model - is highly relevant to the social and political implications the systems might have.

Algorithmic decision-making and corruption

Algorithmic systems, as well as tools and applications based on them, play a growing role in governance and policy-making, which at root deal with questions of power. Naturally then, the deployment of algorithmic systems cannot be isolated from issues that could arise from the abuse of this power, such as corruption. The extensive literature on corruption demonstrates that it tends to change its form according to the system and processes of governance at hand. Just as anti-corruption policies can alter power relations and the nature of corruption in a polity (Rothstein 2011), so too could the use of algorithms in the context of governance and policy making change the forms that corruption takes and even offer new opportunities for venality and the abuse of power.

The problem with algorithmic systems that this paper will highlight is that algorithms in themselves are often assumed to be value-free, impartial and neutral. In practice, algorithms are always shaped by the humans that created them as well as the underlying data that they draw on. As such, they can perpetuate biases, inequalities and also corruption (O'Neil 2016).

Simply digitalising a process does not expunge it of fundamental biases or vulnerabilities to misuse. For instance, transforming a human-driven, paper-based corruption risk assessment into a digital equivalent that employs machine learning to identify potential red flags based on historical data will suffer from many of the same weaknesses as its forebear.

Corruption risks in the application of algorithmic systems can therefore arise both from the discretion and bias of those who design them as well as from the input data with which they are trained.

At the same time, algorithmic systems can also be consciously used for corrupt purposes. Algorithms used to tabulate election results could, for instance, be programmed to favour specific candidates. Equally, algorithmic systems could be used to aid money laundering by complicating money trails and dividing up funds into a complex set of shell companies.

Crucially, algorithmic biases, both intended and unintended, can be hard to detect or comprehend. For this reason, some commentators have labelled algorithms "weapons of math destruction" (O'Neil 2016).

Algorithmic systems as anti-corruption tools

Despite the potential for systemic biases and abuse, it is important to highlight that the capabilities of algorithmic systems also have important implications for *countering* inefficiencies and corruption. Algorithmic systems are able to comb through datasets too large for humans to meaningfully tackle, and can reveal or predict patterns of fraud and corruption that might have otherwise gone undetected. Similarly, discretionary processes previously exposed to high risks of corruption can be made to result in more consistent outcomes through the use of algorithmic systems (Aarvik 2019).

Aarvik (2019), for instance, examines pilot projects that use artificial intelligence to identify corruption risks in public procurement. He cites the example of Mexico, where research institutions have begun to use automated queries to access and analyse past procurement processes to identify corruption risks in specific tenders. Equally, he highlights the example of Ukraine, where Transparency International Ukraine launched a tool to identify and report suspicious tenders using the ProZorro system.

Perhaps most notably, algorithmic systems are already regularly used in the context of anti-money laundering, where they are employed to analyse massive datasets of financial transactions to spot irregularities. As such, they can flag specific transactions to be investigated further or even restrict transactions before they take place (Breslow et al., 2017). The further development of these algorithms is supported by banks, regulators and researchers alike and has yet to reach its full potential (Jullum et al. 2020).

Yet, here too bias also can cause challenges. Algorithms are only as good as the data they are based on. If, for instance, a bank is keeping faulty records or money trails are well hidden, algorithms could learn to recognise illicit financial flows as legitimate (Rainie and Anderson 2017; OECD 2019a).

Bias and opacity in algorithmic decision-making

The capability of algorithmic systems to efficiently process large amounts of data and draw conclusions from that data offers opportunities to many actors in industry and academia, but also to government and society at large (Nolan 2018).

However, as employing algorithmic systems is a means rather than an end in itself, the impact of algorithms in public administration greatly depends on the way a system is designed and developed, and the context in which it is used. Scholars have been pointing to a growing amount of evidence that algorithmic systems can re(produce) and reinforce existing human biases (MIT Technology Review 2020).

Biases that affect the outcomes of algorithmic systems can have different origins. First, the software developers or engineers responsible for designing an algorithm may take decisions that shape outcomes in a certain way. This can happen consciously, for example when an algorithm implements affirmative action principles. However, the design of algorithms can also mirror the implicit and unconscious biases that the developers may hold.

Even when software developers take great care to minimise the risk that their own prejudices interfere with the design of an algorithmic system, the data used to train an algorithm can be another significant source of bias (Barocas and Selbst 2015; Sweeney 2013). For instance, predictive algorithms used in law enforcement attempt to calculate the likely crime rate in different neighbourhoods on the basis of historic arrest data, which may simply serve to encode patterns of racist policing (MIT Technology Review 2020).

In another example, Buolamwini and Gebru (2018) explore the discriminatory effects of insufficiently diverse training data by studying the capability of commercial automated facial analysis algorithms to determine the gender of a person based on a picture of their face. Studying at the gender classification products of three companies (IBM, Microsoft and Chinese competitor Face++), they find that the three companies achieve relatively high overall accuracy for assigning the correct gender to a face.

On closer inspection, however, it becomes apparent that there are significant differences in the error rates between different groups. All three companies perform much better on images of lighter-skinned individuals; in the case of IBM, the difference in error rates is as high as 19%. Buolamwini and Gebru found that the error rate is even higher for darker skinned women: IBM's product, which had the highest error rate, was almost 35% worse at recognising a darker skinned woman than a lighter skinned man. Similarly, in 93% of cases in which Microsoft's software misdiagnosed a person's gender, that person was darker-skinned.

This case points to two important factors responsible for the potentially disparate outcomes of algorithms. On the one hand, the training data with which these commercially available algorithms were trained did not contain the necessary diversity to make sure that the algorithms could robustly recognise gender attributes across different skin shades. On the other hand, at the time of evaluation, none of the companies reported how well their computer vision products performed across gender, skin type, ethnicity, age or other attributes. This oversight should not be attributed to malicious design and development processes, but rather speaks to the lack of diversity among software engineers, as well as a lack of information made publicly available (Stack Overflow 2020).

As these examples demonstrate, algorithms are far from neutral pieces of technology, but reflect the (un)concious preferences, priorities and prejudices of those that build them. Such insights are equally pertinent to the anti-corruption field. Where algorithms are trained to detect incidences of corruption based on historical datasets, the true accuracy of these tools is partly a function of the impartiality of authorities in sanctioning corrupt practices in the past. In settings with a weak rule of law in which anti-corruption campaigns are primarily used by incumbents to target political opponents, the deployment of algorithms in anti-corruption may chiefly result in the more efficient suppression of government critics. Under such conditions, algorithms can potential serve as an instrument of political control.

Beyond being potentially biased, algorithmic decision-making systems can be notoriously opaque, making it difficult and at times impossible to understand how they arrived at an outcome (Pasquale 2015). This is due to the inherent invisibility of the inner workings of algorithms, which is particularly true for algorithmic systems relying on forms of machine learning. As machine learning algorithms do not operate according to a "if → then" logic, but are discovering patterns in the underlying data in a more or less unsupervised manner, it can be nearly impossible to subsequently trace how an algorithmic system produced a given output.

Because of this inherent opacity, algorithmic systems have been compared to "black boxes", and the ethics of using algorithmic systems in public policy and governance has been questioned (Pasquale 2015).

Examples from public administration

There is a clear accelerating trend in the use of algorithmic decision-making systems in the management and delivery of public services (Veale and Brass 2019). However, since many administrative tasks are not straightforward and include 'political' deliberations, there is no consensus regarding the extent to which algorithms can gainfully be deployed in public service, with many scholars and experts arguing that some tasks simply cannot be automated (Lipsky 2010).

Perhaps unsurprisingly therefore, the uptake and implementation of algorithms in public administration and governance differs greatly across countries. In Austria, young parents do not have to apply to receive child benefits; instead, the responsible agency automatically receives and processes the relevant information after the birth of the child, and automatically deposits the funds (Crysmann 2020). In an attempt to create more equal opportunities, an algorithmic system in Belgium helps to assign students to secondary schools. Prior to the implementation of the algorithm, schools would sign up students based on a "first come, first serve" rule, leading to parents camping in front of their preferred school for days to ensure their child would be able to register. Recognising that this disadvantaged some students, the algorithmic system was intended to remedy this situation (Vervloesem 2020). In Norway, machine-learning algorithms are used to interpret x-ray data collected by the Norwegian customs and border control authorities (Deutscher Bundestag 2019).

While algorithmic decision-making systems are already used effectively in many countries' public administrations, they can run into the same structural and institutional pitfalls commercial products regularly do. One example of public administration algorithms gone astray refers to the machine-learning powered system used by the Austrian Public Employment Service to help ascertain the chances of job seekers finding work, and to determine for whom re-education and additional training would be worthwhile. The system, based on historical data of unemployed people navigating the job search, was found to routinely discriminate against women by predicting a lower likelihood of women finding a job, and thus not qualifying for additional (re-)training. In cases of women with a migration background, or with childcare duties, additional points were subtracted, leading to a cumulative disadvantage (Köver 2019). Based on women's historically lower levels of employment, the system thus perpetuated existing social inequalities.

Scientists who scrutinised the software after these flaws became public noted that the system was overall too opaque to understand its error rate. Overall, the Austrian Employment service did not share sufficient amounts of training and testing data, while the models used in the algorithm did not allow external auditors to understand how the system arrived at its output (Cech et al. 2019). This example not only highlights the problem of re-producing societal biases based on flawed historical training data, but also shows the importance of institutional safeguards and protocols to enable independent audits of algorithms, which this paper discusses in the following section on technological tools to achieve greater algorithmic transparency.

Another example refers to COMPAS, a risk assessment software used in courts across the United States to help determine the risk that a suspect or convicted person will re-offend. COMPAS is not the only such software used in US courtrooms, but its alleged systematic bias towards African Americans brought into sharp relief the larger debate on the appropriateness of using such tools in the criminal justice system. The algorithms employed by COMPAS were found to disproportionately predict that black defendants would be more likely to commit a crime in the future (Yong 2018). Additionally, COMPAS was marked by a high error rate, only being able to correctly identify re-offenders in 65% of the cases. This error rate was found to correlate with the success of random survey participants correctly guessing the likelihood that a defendant would re-offend (Dressel and Farid 2018). As such, the algorithm implemented to avoid human biases in the criminal justice system, did not perform significantly better than a group of random study participants recruited from the internet.

These tensions and ambivalences form the backdrop against which governments are increasingly turning to algorithmic systems to streamline governance processes. As algorithmic systems are unquestionably better than humans at processing and analysing large amounts of data, there is a legitimate interest on the part of policy makers to utilise them to make the administration of services more efficient and consistent.

The remainder of the paper considers how to mitigate risks in the use of algorithms in public life that arise from the kind of structural and institutional factors described above, notably the unconscious biases of a system's creators or the quality of the data used for training an algorithmic system.

Given that the opacity of such automated decision-making systems is part of the problem when seeking to identify and resolve biases and potential corruption, it is often assumed that simply increasing transparency in the development and use of algorithms will overcome these issues. The next section of the paper considers the relative advantages and shortcomings of simply providing "more transparency" in relation to the use of algorithms in decision-making processes.

Transparency as antidote?

As discussed above, algorithmic decision-making can potentially entail corruption risks due to the discretion of those who design these systems and the data with which the algorithms are trained. Both of these elements can result in systems that do not serve the common good, but rather perpetuate generational inequalities, prejudice, private gain and other exclusionary practices.

Unsurprisingly therefore, the propensity of algorithms to reproduce patterns of prejudice and bias, and their increased prominence in organisational decision-making, have prompted calls for better transparency and increased accountability of algorithmic decision-making systems (Diakopoulos 2015; Pasquale 2015).

Being able to *'see'* how an algorithm works is assumed to allow for greater oversight in its application, as well as to ensure fair and non-discriminatory outcomes.

Algorithmic transparency refers to the principle that the factors that influence the decision of an algorithmic system should be transparent - or visible - to the people employing or affected by the outcomes of the algorithmic system (Diakopoulos and Koliska 2017). Although algorithmic transparency and algorithmic accountability are often used interchangeably, the underlying concepts differ.

Whereas algorithmic transparency suggests that the inputs into a system and the workings of that system must be known, algorithmic transparency does not necessarily require the outcome of the system to be fair. Algorithmic accountability refers to the notion that the institutions building or employing algorithms must be accountable for the outcomes of decisions made by algorithmic systems (Diakopoulos 2015). Understanding the output of an algorithmic system in order to be able to hold its owners and designers to account presupposes a large degree of algorithmic transparency.

Calls for greater transparency in the design and application of algorithms seem to be an intuitive response to the problems identified. Indeed, the view that transparency is instrumental to successful efforts to reduce corruption is an idea that is as old as the fight against corruption itself, and one encapsulated in the very name Transparency International.

On a conceptual level, this idea is rooted in the principal-agent perspective on the control of corruption, which contends that corruption is rooted in the information asymmetries between citizens (the principals) and civil servants (the agents). The lack of information on the agents' work makes it impossible for principals to control them, which leads to greater potential for unrestrained discretionary behaviour than can create opportunities for corruption, such as self-enrichment (Klitgaard 1991; Groenendijk 1997; Meijer 2013). It follows from this that greater transparency should, generally speaking, lead to better governance and greater accountability.

While this link seems obvious in theory, it is less well established in practice. Transparency *can* lead to greater accountability, whether or not this is the case strongly depends on what kind of information is provided and how (Zúñiga 2019). Lindstedt and Naurin (2010) point to contextual factors such as media freedom or education as key in ensuring higher levels of transparency results in greater accountability and less corruption.

Where these kind of necessary enabling conditions are absent, greater transparency alone is unlikely to significantly increase accountability. Bauhr and Grimes (2014) even find that in highly corrupt settings without a functioning legal system, transparency can lead to resignation and frustration in the population. They argue that increased knowledge about abuses of power such as corruption in settings with little accountability is likely to lead to disillusionment among citizens (see also Mungiu-Pippidi 2013; Persson, Rothstein & Teorell 2011).

This matters for several reasons. First, while there is a natural inclination to demand greater algorithmic transparency, the literature indicates there is no clear linear relationship between transparency, accountability and a better control of corruption. Sunlight may be the best disinfectant, but algorithmic transparency alone is not a straightforward antidote to the thorny issues related to the growing use of algorithms in public administration.

Second, it can be particularly challenging to make algorithms "transparent" in a meaningful and comprehensible way. To ensure transparency leads to greater accountability, the data made available should allow the scrutiny of potential bias built in the algorithm or in the data with which it is trained. Yet algorithms are often opaque if not impenetrable. While making rule-based "if → then" algorithms readable to humans is generally possible, this is incredibly difficult in the case of machine-learning algorithms given that they tend to operate as a kind of black box: their outputs can be impossible to interpret or trace back to the original inputs. For machine-learning systems, therefore, transparency is particularly essential at the design stage, so those affected are aware of the values programmed into the system and the nature of the training data.

Third, it is worth noting that there is currently no consensus as to what algorithmic transparency consists of. The many frameworks, ethical guidelines and principles for the design and development of algorithms that have emerged over the past years highlight various dimensions of transparency. Algorithmic transparency can refer, for instance, to measures intended to ensure quality and fairness during the design and development stages, the type, source and quality of input or training data, the algorithmic methods and models used, or the testing, evaluation and monitoring processes used during deployment and the handling of their results (Zweig 2019). An overview of the various possible components of algorithmic transparency is provided in Table 1 below.

Ultimately, when talking about algorithmic transparency as a means to greater accountability, there is a need to understand what type of transparency is meant and how it can lead to accountability. The next section of this paper considers this to offer perspectives as to how algorithmic transparency can, and cannot, help to address issues related to potential shortcomings in algorithmic systems.

The promise and peril of algorithmic transparency

How, then, can algorithmic transparency be meaningfully achieved? Firstly, it is helpful to understand that there is no exact definition of what makes an algorithmic system transparent. Transparency is not a black and white notion (i.e. a system is either transparent or not). Instead, algorithmic systems can come with different degrees of transparency, depending on a combination of their technical properties and governance processes.

In fact, depending on the 'audience' i.e. whether a system is intended to be transparent for the general public, regulators or researchers, different degrees of transparency are necessary and useful. It is helpful to distinguish between two goals of algorithmic transparency and accountability. Firstly, to understand the general process according to which an algorithmic system works. Secondly, to understand how an algorithmic system arrives at an individual outcome.

Technological tools to reduce opacity

When trying to understand the workings of an algorithmic system, it is not very informative to scrutinise the actual steps an algorithm took. The answer to such a question could be a simplistic explanation of the general process, which does not yield much information on why the system performs certain tasks. The answer could also be a complete list of steps taken, in the form of the complete algorithm or model, which will likely be so complex that not even experts would understand why each step was taken. Both answers would likely be useless for citizens, regulators, and most experts.

Nonetheless, this is not in itself an argument against disclosure and transparency. In other domains, there is no expectation that the average citizen will fully comprehend the technical details; most people are not expected to understand the legal arguments in published judgements. Yet court decisions are still expected to be published so that justice is seen to be done, and experts such as lawyers and journalists can interpret them on behalf of the public. That being said, simply disclosing the steps an algorithm takes to reach an output is generally not a very meaningful approach to achieving transparency (Koene et al. 2019).

Fortunately, several technical methods can help reduce the opacity of algorithmic systems. One tool is the analysis of input data. Analysing input data is a process during which all information made available to a system for making a decision is scrutinised to ensure that the input data is not flawed, incomplete, or biased. Input data analysis usually requires proprietary information and may thus be difficult to achieve without cooperation from the organisation that built the algorithmic system. Input data analysis does not give direct insight into how a system 'works' but can be a useful step towards understanding a system and its components (Eeckhout et al. 2003).

Another tool is the statistical analysis of outcomes. This can, for example, include the use test data sets to determine whether a system behaves robustly across different populations and does not lead to unforeseen or unwanted outcomes. This does not necessarily shed light on why a system has taken which step, but can help to understand whether there are larger issues that warrant closer inspection. Outcome analysis can be especially useful in the context of statistical systems, though it can be more difficult to achieve in learning systems (Datta et al. 2016).

Another related method is blackbox or sensitivity testing. The goal of this approach is to better understand how a system reacts by providing many inputs or test cases that show very slight differences. For example, by sending hundreds of only slightly different credit applications to an automated credit scoring system, researchers can infer information about the system from analysing its outputs (Zweig 2019). This is an especially important method in the context of machine learning algorithms, where even researchers and the developers of a given system may only have limited insight into how it produces specific outcomes (Saltelli et al. 204).

Challenges to technological approaches to algorithmic transparency

Although there are technical tools available to reduce the opacity of algorithmic systems ex post, there are also several technical challenges that can make systems more or less transparent. One such challenge is the complexity of algorithmic systems: while one module may be understandable, the interactions of multiple parts of the system may be too complex to effectively scrutinise. Another challenge relates to the relationships between decisions - as some algorithms are often used to solve several problems at once, the solutions to those problems depend on each other. Explaining one particular decision may thus be challenging or impossible. Finally, machine learning algorithms, which build a model based on input data, are routinely so complex that they are often opaque even to their developers. This is exacerbated in the case of systems that learn iteratively and therefore continually change (Koene et al. 2019).

Even when algorithmic systems rely on methods that could allow for post-deployment auditing, transparency can be difficult to achieve. Most commercial software providers refuse to disclose the exact working of their algorithms, or the input data used, on the basis that this is part of their intellectual property rights. At heart a commercial product - and often difficult to separate from its input or training data - algorithmic systems are usually patented or protected under intellectual property and copyright law. This commercial secrecy can pose many legal hurdles to “opening up” algorithmic systems, especially where governments procure algorithmic systems from private businesses.

While the issue of commercial patents is complex, there is a case to be made that any algorithm used by a public body should be able to be reviewed and audited to ensure it is serving the public good. Furthermore, given that companies want better data to train their algorithms, greater transparency could lead to a win-win, in which users find and correct erroneous data that affects them and companies benefit from more accurate training data (Tow Center 2015).

Nonetheless, depending on jurisdiction, legal protections related to intellectual property can be very difficult to circumvent, as exemplified by the case of a Californian defendant who was sentenced to life without parole based on the output of a DNA analysis software. When an expert witness sought to review the source code of the programme, the vendor of the software claimed that it was a trade secret, an argument upheld by the Court (Wexler 2018).

Beyond legal and technical barriers, making an algorithm completely ‘transparent’ may also interfere with the rights of others, such as the right to privacy, and could make the algorithm vulnerable to manipulation or ‘gaming’ (Ananny and Crawford 2018; Wachter, Mittelstadt and Russell 2017).

Yet another dimension of complexity to the question of algorithmic transparency refers to what actually makes an algorithmic system transparent, accountable or explainable, and for whom. Most consumers interacting with algorithms who are curious about the processes behind the outcomes they are presented with would not be served well if given access to the source code of the respective software.

For instance, a user might want to understand why their Google search results were ordered in the way they were. Inviting that user to go through the millions of data points feeding into such a ‘decision’ would most likely not help them understand how a certain result came to be. While promising alternatives, like explaining outcomes by providing counterfactuals, are being explored, the question of what constitutes explainability to whom should not be underestimated (Wachter, Mittelstadt and Russell 2018).

Beyond these more practical difficulties surrounding transparency, Annany and Crawford (2018) point to more general limitations. They argue that algorithms are not just technical objects, but always a combination of code and human norms, behaviours, practices and relationships, and must therefore be understood as a system. Algorithmic transparency does therefore not just necessitate an interrogation of an algorithmic system itself, but also of the context it was produced in, its intended purpose, as well as the rules that govern how it is made, used and evaluated.

Guiding principles for algorithmic transparency

Building and deploying algorithmic systems are complex tasks, and the many decisions taken along the way - from the design of a model, to the collection and use of training data to the context in which it is used and by whom - can have significant implications for the transparency, and by extension the accountability of a system. In the case of algorithmic systems built for or by public administrations, it is especially important to move towards improved transparency and accountability to help avoid unfair outcomes, bias and corruption caused or facilitated by algorithms.

In recent years, the number of available frameworks, ethical guidelines and principles for the design and development of algorithms have proliferated. Companies involved in the development of algorithmic systems, like Google or Microsoft, have published guidelines for the development of ethical algorithms or artificial intelligence. In the European Union, the European Commission will soon propose regulation to embed the principles of non-discrimination, human oversight, safety, privacy and accountability into algorithmic systems (European Commission 2020a).

Perhaps most significantly, the OECD's (2019b) *Recommendation of the Council on Artificial Intelligence* contains a specific article on transparency and explainability that can be seen as an overarching framework for the sector. This calls on all relevant stakeholders to "commit to transparency and responsible disclosure regarding AI systems." More specifically, it outlines some of the key elements of an algorithmic transparency and accountability policy. The *Recommendation* appeals to those involved in developing and using artificial intelligence including algorithm-driven automated decision-making systems to:

1. "foster a general understanding of AI systems;

2. make stakeholders aware of their interactions with AI systems, including in the workplace;

3. enable those affected by an AI system to understand the outcome; and,

4. enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."

As algorithmic systems deployed in the public sector can have wide reaching impacts, some organisations have begun offering specific guidance to public institutions to help embed transparency and accountability into algorithmic systems they may develop, procure, or use.

For instance, the Association for Computing Machinery (2017) sets out seven key principles for algorithmic transparency and accountability.

1. "Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

2. Access and redress: Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

3. Accountability: Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

4. Explanation: Systems and institutions that use algorithmic decision-making are encouraged to produce explanations

regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.

5. Data Provenance: A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorised individuals.

6. Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

7. Validation and Testing: Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public."

In addition, an interactive tool from the Center for Democracy and Technology (2017) offers insight on questions to ask throughout the process of designing, building, testing and implementing an algorithm to help prevent bias and discrimination.

Similarly, the AI Now Institute of New York University has published an Algorithmic Accountability Policy Toolkit that answers some of the most common questions regarding algorithms and automated systems and collects various resources to help practitioners to navigate algorithmic accountability. The toolkit includes background information on algorithmic systems and artificial intelligence technologies and provides an overview of how algorithms are used in public administrations, as well as covering the issue of potential sources of bias. The toolkit thus aims to sensitise practitioners to potential pitfalls of deploying algorithmic systems in the public sector rather than prescribing a set of norms or protocols (AI Now Institute 2018).

Another excellent resource is the Algo.Rules project by the German think tanks Bertelsmann Stiftung and iRights.Lab. The project proposes nine principles to help researchers, designers, developers and users of algorithmic systems take decisions during the planning, development and deployment of algorithmic systems that allow for the review and auditing of algorithmic systems (Puntschuh and Fetic 2020). While non-exhaustive, the nine principles offer a good basis to enhance algorithmic transparency throughout the lifecycle of an algorithmic system deployed by or for the public sector. They are 1) strengthen competency, 2) define responsibilities, 3) document goals and anticipated impact, 4) guarantee security, 5) provide labelling, 6) ensure intelligibility, 7) safeguard manageability, 8) monitor impact and 9) establish complaint mechanisms.

Differentiating between the planning, development and deployment phases of algorithmic systems, the Algo.Rules project proposes key steps for every phase, through which the nine principles can be operationalised.

In the planning phase, those involved in the planning of an algorithmic system should:

1) identify their needs and formulate the goals the system should accomplish;

2) estimate the impact of the system;

3) identify and involve affected stakeholders to take their concerns and needs into account;

4) identify and apply relevant regulatory frameworks;

5) outline the desired system and draw up a development plan.

Heeding these steps should ensure the involved parties to be aware of their needs, how they can be technologically met, and what potential roadblocks, sensitivities or local factors they should be aware of (Puntschuh and Fetic 2020).

During the development stage, Algo.Rules suggests that developers should, in a first step, agree on set design requirements - for example regarding the system's functionalities, safety features, the labelling of the system's outcomes, as well as its explainability and evaluation. This is the stage in which key technological features are determined that can help to make the algorithmic system accountable after its deployment. In practice, this could mean that the software application in which the algorithmic system is embedded includes a complaint or appeals function.

During the development stage, it is also important to document the development work, components used and decisions taken during that process (Puntschuh and Fetic 2020). One example of how documentation can be help to achieve greater transparency are datasheets that record the nature and provenance of datasets used in the development or training of an algorithmic system. Taking inspiration from the datasheets that accompany components in the electronics industry, Gebru et al (2018) propose to create similar information repositories to shed light on the motivation, composition or collection process behind the development of algorithms.

In the deployment stage, Algo.Rules propose a catalogue of specific best practices to ensure transparency and accountability. Among them, Algo.Rules proposes that algorithmic systems used in the public sector should always be labelled as such and offer information on how results are achieved. This entails providing information on the factors that the system's decision-making process considers, as well as information about any weighting of factors in terms of importance, and also the objectives and limitations of the system. Algo.Rules also highlight the need to provide specialised training to those public servants using the system to enable them to understand its functionalities and spot possible inaccuracies.

Finally, the outcomes created by algorithmic systems used in the public sector should be constantly documented and evaluated to make sure that the system works according to the intended goals and established benchmarks (Puntschuh and Fetic 2020). A specific example of this is the work of researchers at Google, who have been developing a framework to enable better monitoring and evaluation of an algorithmic system called model cards. Model cards are documents that detail how machine-learning models should perform in a variety of conditions, such as different cultural contexts, different skin tones, geographic location or other categories. Model cards also explain in which context a model is intended to be used, and how its performance can be evaluated (Mitchell et al. 2019).

The Algo.Rules offer one approach to achieve algorithmic transparency in algorithmic systems deployed by or for the public sector. While widely applicable, organisations developing, deploying, and ultimately governing algorithmic systems will need to reckon with their particular context, stakeholders, needs and goals and tailor their approach to algorithmic transparency accordingly.

There are nonetheless certain common types of data that are likely to be relevant across most contexts. The Tow Center for Digital Journalism (2015) has set out specific types of data that owners of algorithmic systems could disclosure in order to enhance transparency, grouped into five categories.

Table 1: possible components of algorithmic transparency (taken from Tow Center (2015))

Category	**Type of Data to disclose**
Human Involvement	The **goal, purpose, and intent** of the algorithm **Who** is the developer, **who** has direct control over the algorithm, **who** has oversight and is accountable
Data	The **quality** of the data, including its accuracy, completeness, and uncertainty, as well as its timeliness, magnitude and assumptions Dimensions of **data processing** such as how data was collected, transformed, vetted, and edited Whether the data the algorithms uses is **private or public**
The Model	The **features** or **variables** used in the algorithm, as well as any **weighting** of these factors and the **rationale** for this weighting The **assumptions** (statistical or otherwise) behind the model
Inferencing	**Benchmarks** of the inferences used against standard datasets The margin of **error** and **accuracy** rate (number false positives & false negatives) Steps are taken to remediate known errors Range of those confidence values as a measure of **uncertainty** in the outcomes.
Personalisation	Types of **personal information** being used

Achieving algorithmic transparency and accountability in a meaningful sense cannot, therefore, be achieved with the flick of a switch. It requires the constant and continuous documentation, monitoring, and evaluation of algorithmic systems at every stage of their life cycle. To empower citizens confronted with ever-more complex algorithmic systems in their interactions with the state, it is equally important to invest in digital literacy efforts to enable people to understand how algorithms work and what their limitations are.

A final word of caution on the complex inter-relationships between algorithms, transparency, accountability and fairness. No matter how transparent or accountable algorithmic systems are, their results can still be perceived as deeply unfair. For example, think of the algorithm that helps determine where Belgian children go to school that was mentioned above. While some might find it the fairest option to have an algorithm decide this question, others' subjective view of fairness might differ. While the current iteration of the algorithm does take student preferences into account, not all preferences are held equally strongly or as well founded (Verfloesem 2020).

This example should serve as a reminder that technological 'solutions' to complicated questions - like the distribution of a scarce resources - cannot resolve underlying social or political tensions, and, by extension, algorithms cannot simply replace the role of human political institutions in mediating these tensions (Morozov 2014a).

Conclusion

Achieving algorithmic transparency and accountability is a complex matter, and in order to offer meaningful insights to and empowerment of citizens, algorithmic transparency must not be an afterthought.

Rather, it is a goal that must be considered from the outset. Specific protocols to achieve this objective need to be adopted and implemented during the process of planning, designing, testing, documenting, deploying, auditing and evaluating algorithms. Moreover, to achieve real algorithmic transparency, the entire context in which code is written, tested and used has to be taken into account during this process.

This is no simple task. Yet the work of Algo.Rules, the Center for Democracy and Technology and other similar organisations illustrates that there are suitable methods to make algorithmic systems more understandable and auditable. Especially in the public sector, great care should be taken to create software that lends itself to transparency, accountability and explainability, both before and after deployment.

Policy-makers contemplating the adoption of algorithmic systems in public administration should consider whether measures have been taken to enhance algorithmic transparency and comprehensibility during the procurement or development of these algorithms. Clearly, much work remains to be done, and in the meantime, algorithms used in the public sector should employ statistical methods that lend themselves to explainability.

Just as important is the fact that not every problem or administrative process lends itself to automation. Many government processes entail the weighing of political or normative factors and are as such ill-suited to algorithmic decision-making. To avoid undesirable, opaque and unfair outcomes resulting from the use of algorithms in public administration, careful evaluations of the problem at hand are necessary before deciding whether to adopt an algorithmic system. Another important aspect is data quality and data management. For software that is used in the public sector, training and testing data must be of high quality, suitable for the task at hand, findable, interoperable, accessible and reusable. Using the wrong set of data for training and testing can result in algorithmic bias and increase the opacity of the algorithmic system.

While the importance of algorithmic systems in public administration is only going to increase in the coming years, algorithmic decision making systems also play an increasingly significant role in consumer applications and on online platforms. Algorithmic decision making systems determine how internet users find and access information, content, services, and goods online, and can thus have a massive impact on how people's political, economic and social lives.

This increased reliance on algorithmic systems and the need to improve transparency in their use is also mirrored in forthcoming EU legislation. The Digital Services Act package, proposed by the European Commission in December 2020, suggests new rules to require large commercial platforms to be transparent about the main parameters that shape algorithmic decision-making systems used to curate content and moderate content. The legislative initiative also proposes new obligations for intermediaries to be more transparent when, and for what purpose algorithmic systems are used on their platforms, and to which extent such systems are under the control of human review in the context of content moderation (European Commission, 2020b).

While it remains to be seen how the proposed rules will develop during the legislative negotiations, the package - in addition to the European Commission's white paper on artificial intelligence published in 2020 - clearly marks the intent of the European Union to embed the principles of transparency and fairness in algorithmic systems.

Ultimately, it is to be hoped that regulators around the world are able to devise and enforce safeguards that strike an appropriate balance between realising the benefits of algorithmic decision-making while minimising potential harm to citizens and consumers.

References

Adam, I., and Fazekas, M. 2018. “Are emerging technologies helping win the fight against corruption in developing countries?” Pathways for Prosperity Commission Background Paper Series; no. 21. Oxford, United Kingdom. http://www.govtransparency.eu/wp-content/uploads/2019/02/ICT-corruption-24Feb19_FINAL.pdf

AI Now Institute. 2018. Algorithmic Accountability Policy Toolkit. https://ainowinstitute.org/aap-toolkit.html.

Ananny, M., & Crawford, K. 2018. “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability”, *new media & society*, *20*(3), 973-989.

Apolitical. 2020. *The digital government atlas 2.0: the world's best tools and resources.* https://apolitical.co/en/solution_article/the-digital-government-atlas-the-worlds-best-tools-and-resources.

Artificial Solutions. 2020. *Why most chatbots are annoying.* https://www.artificial-solutions.com/blog/why-chatbots-are-annoying-make-sure-yours-isnt

Association for Computing Machinery. 2017. “Principles for Algorithmic Transparency and Accountability.” https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

Athey, S. 2017. “Beyond prediction: Using big data for policy problems”. *Science*, *355*(6324), 483-485.

Barocas, S, Selbst, A.D. 2015. *Big data's disparate impact.* Rochester, NY: Social Science Research Network

Breslow, S., Hagstroem, M., Mikkelsen, D., & Robu, K. 2017. *The new frontier in anti–money laundering*. *https://www.mckinsey.de/~/media/McKinsey/Business%20Functions/Risk/Our%20Insights/The%20new%20frontier%20in%20anti%20money%20laundering/The-new-frontier-in-anti-money-laundering.pdf*

Buolamwini, J., & Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Cech, F., Fischer, F., Human, S., Lopez, P. & Wagner, B. 2019. Dem AMS-Algorithmus fehlt der Beipackzettel. *Futurezone.* https://futurezone.at/meinung/dem-ams-algorithmus-fehlt-der-beipackzettel/400636022

Center for Democracy and Technology. 2017. “So, you want to build an algorithm”. https://www.cdt.info/ddtool/

Coglianese, C. 2016. "Robot regulators could eliminate human error", *San Francisco Chronicle*. https://www.sfchronicle.com/opinion/article/Robot-regulators-could-eliminate-human-error-7396749.php

Crysmann, T. 2020. “Gefährliche Rechenspiele”, *Süddeutsche Zeitung,* https://www.sueddeutsche.de/digital/werkstatt-demokratie-algorithmen-verwaltung-diskriminierung-datenschutz-1.5127319

Datta, A., Sen, S., & Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)* (pp. 598-617). IEEE.

Diakopoulos, N. 2015. Algorithmic Accountability. *Digital Journalism,* 3 (3), 398–415.

Diakopoulos, N., & Koliska, M. 2017. Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809-828.

Domingos, P. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Dressel, J., & Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Eeckhout, L., Vandierendonck, H., & De Bosschere, K. 2003. Quantifying the impact of input data sets on program behavior and its applications. *Journal of Instruction-Level Parallelism*, *5*(1), 1-33.

European Commission. 2020a. White Paper on Artificial Intelligence - A European approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. 2020b. The Digital Services Act package. https://ec.europa.eu/digital-single-market/en/digital-services-act-package

Fox, J. 2007. The uncertain relationship between transparency and accountability. *Development in practice*, *17*(4-5), 663-671.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Groenendijk, N. 1997. A principal-agent model of corruption. *Crime, Law and Social Change*, *27*(3-4), 207-229.

Höchtl, J., Parycek, P., & Schöllhammer, R. 2016. Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, *26*(1-2), 147-169.

Jullum, M., Løland, A., Huseby, R.B., Ånonsen, G. and Lorentzen, J. 2020, "Detecting money laundering transactions with machine learning", *Journal of Money Laundering Control*, Vol. 23 No. 1, pp. 173-186.

Klitgaard, R. 1991. *Controlling corruption*. University of California Press.

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. 2019. *A governance framework for algorithmic accountability and transparency.* Panel for the Future of Science and Technology, European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262t/EPRS_STU(2019)624262_EN.pdf

Köver, C. 2019. Streit um den AMS-Algorithmus geht in die nächste Runde. *Netzpolitik.org.* https://netzpolitik.org/2019/streit-um-den-ams-algorithmus-geht-in-die-naechste-runde/

Kroll, J. A. Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G. & Yu, H. 2017. ‘Accountable algorithms’, University of Pennsylvania Law Review, Vol. 3, p. 633

Lindstedt, C., & Naurin, D. 2010. Transparency is not enough: Making transparency effective in reducing corruption. *International political science review*, *31*(3), 301-322.

Lipsky, M. 2010. Street-level bureaucracy: Dilemmas of the individual in public services.New York: Russell Sage Foundation.

Loeber, K. 2018. Big Data, Algorithmic Regulation, and the History of the Cybersyn Project in Chile, 1971–1973. *Social Sciences*, *7*(4), 65.

Meijer, A. 2013. Understanding the complex dynamics of transparency. *Public administration review*, *73*(3), 429-439.

Merriam-Webster Dictionary (2020). *Artificial Intelligence.* https://www.merriam-webster.com/dictionary/artificial%20intelligence

MIT Technology Review. 2020. *Predictive policing algorithms are racist. They need to be dismantled.* https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. 2016. The ethics of algorithms: Mapping the debate. Big Data & Society , 3 (2), 2053951716679679.

Mohabbat Kar, R., Thapa, B. E.P., Parycek, P. (Hg.) (2018). *(Un)Berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft.* Berlin: Kompetenzzentrum Öffentliche IT

Morozov, E. (6 October 2014a). The rise of data and the death of politics. *The Guardian.* https://www.theguardian.com/technology/2014/jul/20/rise-of-data-death-of-politics-evgeny-morozov-algorithmic-regulation

Morozov, E. 2014b. The Planning Machine. *The New Yorker.* Retrieved on 18 December 2020 from https://www.newyorker.com/magazine/2014/10/13/planning-machine

Mungiu-Pippidi, A. 2013. Controlling corruption through collective action. *Journal of Democracy*, *24*(1), 101-115.

Nolan, A. 2018. Artificial intelligence and the technologies of the Next Production Revolution. In: *OECD Science, Technology and Innovation Outlook - Adapting to Technological and Societal Disruption.* https://doi.org/10.1787/sti_in_outlook-2018-en

OECD. 2019a. “Anti-Corruption and Integrity Forum: Tech Topics of the Forum.” https://www.oecd.org/corruption/integrity-forum/tech-topics/

OECD. 2019b. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pasquale, F. 2015. *The black box society*. Harvard University Press.

Persson, A., Rothstein, B., & Teorell, J. 2013. Why anticorruption reforms fail—systemic corruption as a collective action problem. *Governance*, *26*(3), 449-471.

Stack Overflow (2020). *Stack Overflow Developer Survey 2020.*

https://insights.stackoverflow.com/survey/2020#overview

Puntschuh, M. and Fetic, L. 2020. Handreichung für die digitale Verwaltung - Algorithmische Assistenzsysteme gemeinwohlorientiert gestalten. *Bertelsmann Stiftung and iRights.Lab.* https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf

Rainie, L. & Anderson, J. 2017. Code-Dependent: Pros and Cons of the Algorithm Age. *Pew Research.* https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/

Rothstein, B. 2011. Anti-corruption: the indirect 'big bang'approach. *Review of International Political Economy*, 18(2), 228-250.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. 2004. Sensitivity analysis in practice: a guide to assessing scientific models (Vol. 1). New York: Wiley.

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

Tow Center. 2015. Towards a Standard for Algorithmic Transparency in the Media. https://medium.com/tow-center/towards-a-standard-for-algorithmic-transparency-in-the-media-81c7b68c3391.

Veale, M., & Brass, I. 2019. Administration by algorithm? Public management meets public sector machine learning. *Public Management Meets Public Sector Machine Learning*.

Wachter, S., Mittelstadt, B., & Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

Vervloesem, K. 2020. In Flanders, an algorithm attempts to make school choice fairer. *Automating Society Report.* https://algorithmwatch.org/en/project/automating-society/

Wexler, R. 2018. Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stan. L. Rev.*, *70*, 1343.

Yong, E. 2018. A Popular Algorithm Is No Better at Predicting Crimes Than Random People. The Atlantic. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

Zerilli, J., Knott, A., Maclaurin, J. et al*.* 2019. 'Algorithmic Decision-Making and the Control Problem', *Minds & Machines*, vol.29**:** 555–578. https://doi.org/10.1007/s11023-019-09513-7

Zúñiga, N. 2019. Does more transparency improve accountability? U4 Anti-Corruption Helpdesk.

Zweig, K. 2019. *Ein Algorithmus hat kein Taktgefühl: Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. Heyne Verlag.

“Anti-Corruption Helpdesk Answers provide practitioners around the world with rapid on-demand briefings on corruption. Drawing on publicly available information, the briefings present an overview of a particular issue and do not necessarily reflect Transparency International’s official position.”

Transparency International
International Secretariat
Alt-Moabit 96
10559 Berlin
Germany

Phone: +49 - 30 - 34 38 200
Fax: +49 - 30 - 34 70 39 12

tihelpdesk@transparency.org
www.transparency.org

blog.transparency.org
facebook.com/transparencyinternational
twitter.com/anticorruption

Transparency International chapters can use the Helpdesk free.
Email us at ***tihelpdesk@transparency.org***