"Automating Anticorruption?"

Ceva, Emanuela; Jimenez Garcia, Maria Carolina

How to cite

CEVA, Emanuela, JIMENEZ GARCIA, Maria Carolina. 'Automating Anticorruption?'. In: Ethics and Information Technology, 2022, vol. 24, n° 8, p. 48. doi: 10.1007/s10676-022-09670-x

This publication URL: https://archive-ouverte.unige.ch/unige:164975
Publication DOI: 10.1007/s10676-022-09670-x

Ethics and Information Technology (2022) 24:48
https://doi.org/10.1007/s10676-022-09670-x

ORIGINAL PAPER

Automating anticorruption?

Emanuela Ceva[1] · **María Carolina Jiménez**[1]

Accepted: 14 October 2022

Abstract
The paper explores some normative challenges concerning the integration of Machine Learning (ML) algorithms into anticorruption in public institutions. The challenges emerge from the tensions between an approach treating ML algorithms as allies to an exclusively legalistic conception of anticorruption and an approach seeing them within an institutional ethics of office accountability. We explore two main challenges. One concerns the variable opacity of some ML algorithms, which may affect public officeholders' capacity to account for institutional processes relying upon ML techniques. The other pinpoints the risk that automating certain institutional processes may weaken officeholders' direct engagement to take forward-looking responsibility for the working of their institution. We discuss why both challenges matter to see how ML algorithms may enhance (and not hinder) institutional answerability practices.

Keywords Corruption · Anticorruption · Office accountability · Machine learning algorithms · Opacity · Artificial intelligence

Introduction

Is it justifiable to integrate Machine Learning (ML) algorithms into the fight against corruption within public institutions? If so, what role could they have?

In public institutions, corruption may indicate a vice of the officeholders' conduct (e.g., embezzlement or misappropriation) and/or the officeholders' implication in an institutional malpractice (e.g., systemic bribery or clientelism—Miller, 2017; Philp, 1997; Thompson, 2018; Warren, 2004). Accordingly, current studies of corruption in the social sciences and practical philosophy have focused on the individual or collective actions of those who hold a public office and wield public power (Ceva & Ferretti, 2021a; Heywood, 2015; Philp, 1997; Rose-Ackerman, 1999). Anticorruption thus requires the identification of corrupt individual or collective actions, their investigation and, possibly, their punishment or rectification (for an overview, see Mungiu-Pippidi et al., 2015).

Within this general approach to anticorruption, ML algorithms may serve, *inter alia*, to estimate the risk of officeholders' corrupt behavior. High scores in corruption risk indicators can be used as "red flags" triggering investigations leading to the targeted officeholders' prosecution and/or dismissal. Moreover, ML algorithms have been growingly used to automate the making of such public decisions as those concerning the selection of candidates requesting social benefits. This integration intends to curb the risk of corruption by limiting the (ab)use of officeholders' discretion.

Despite the increasing practical relevance of ML algorithms for anticorruption, conceptual frameworks to analyze the *ethical* implications of integrating Artificial Intelligence (AI)-driven technologies into anticorruption strategies are scant (Kobis & Starke, 2020). To start filling this gap, we specify some normative challenges of integrating ML algorithms into public institutional action. To analyze such challenges is important to pinpoint the risks implicated in resorting to ML techniques in the context of an ethics of office for public institutional action. The focus on anticorruption, while specific, bears some general interest too. Corruption is arguably one of the main public dysfunctions. Therefore, the strategies for countering it are indirectly telling of the aspirations to uphold a certain normative ideal of public institutional action. While focused on the ethics of anticorruption, our discussion can also open the path for a broader

✉ Emanuela Ceva
Emanuela.Ceva@unige.ch

[1] Department of Political Science & International Relations, University of Geneva, Geneva, Switzerland

Published online: 09 November 2022

critical reflection about the scope for employing ML algorithms to uphold public institutional action more generally.

We start by illustrating the possible contributions of ML algorithms to anticorruption in public institutions ("The integration of ML algorithms into anticorruption strategies" section). The appraisal of such contributions is indicative of some general limits of a purely legalistic and regulatory approach to anticorruption, which current ML techniques presuppose and implement. Such an approach targets corrupt behavior that entails some formal rule violation, which ML algorithms are capable of identifying or limiting ("ML algorithms and the limits of an exclusively legalistic approach to anticorruption" section). In the "Anticorruption within an institutional ethics of office accountability" section, we introduce an additional (but not alternative) approach. Such an approach targets also ethically relevant instances of corruption in public institutions, which do not necessarily entail formal rule-breaking. Such instances require giving ethical guidance to officeholders to uphold by their action the working of their institution. In the "The normative challenges of integrating ML algorithms and office accountability" section, we discuss some normative challenges of integrating ML algorithms into anticorruption, when this latter is not only understood in legalistic terms but also as a component of an institutional ethics of *office accountability* (2021b; Ceva & Ferretti, 2021a). We then conclude ("Conclusion" section)..

Throughout our discussion, we shall be careful not the throw the baby of ML algorithms out with the bathwater of some of their criticalities. We invite an analysis of such criticalities. Such an analysis is important to appraise how resorting to ML algorithms may favor or undercut the development of an ethics of office capable of resisting corruption and sustaining public institutional action over time.

The integration of ML algorithms into anticorruption strategies

Tech corporations and development agencies actively promote AI amongst governments around the globe as a tool to tackle various institutional dysfunctions, including corruption. In the context of anticorruption, ML algorithms are the most relevant form of AI (Rahwan et al., 2019). ML algorithms are often used as a tool for predictive analytics. Algorithmic outputs may take the form of statistical probabilities with respect to the occurrence of a particular behavior or event. ML algorithms may sustain anticorruption in two main ways. Some of them may facilitate the detection or prediction of some officeholders' corrupt behavior as concerns the use they (can be expected to) make of their power of office. Some other applications enable the automation of decision-making processes that are prone to corruption in order to "protect" such processes from the abuse of human discretion (Aarvik, 2019; Lima & Delen, 2020; López-Iturriaga & Sanz, 2018; Petheram et al., 2019).

The first kind of ML applications draws on governmental records accessible in a digital form, as well as other datasets from a variety of sources. The analyses of such inputs can estimate the risk of officeholders' corrupt behavior in contexts such as government procurements, taxation, or the access to public services. ML algorithms may thus score various offices and flag those at high risk of corruption. Such flags may also work as indicators for launching some additional investigation and, possibly, lead to punitive measures.

An example is the *Prozorro* e-procurement platform, launched in 2016 by the Ukrainian government to detect corrupt tenders (Kovalchuk et al., 2019; Petheram et al., 2019). All public procurements run through the platform and all data are accessible online, including outcomes, successful suppliers, and contract values (Petheram et al., 2019, pp. 10–11). Transparency International Ukraine has also developed *Dozorro*, an ML technology to monitor governmental procurements. Experts labelled 3.500 tenders as risky or non-risky to train the ML application. The results of *Prozorro* and *Dozorro* can be reported to the authorities for formal investigation. Another well-known example comes from Brazil; it was developed by the Office of the Comptroller General to estimate the probability of corrupt behavior amongst its civil servants. The ML algorithm was trained using large data sets containing information of civil servants' past convictions, and is able to weight hundreds of indicators (e.g., criminal records, political affiliations, education, business and shareholder relations) to calculate the probabilities of a civil servant being corrupt (Aarvik, 2019; Marzagão, 2017). The calculus only requires entering someone's social security number into the dashboard. However, Brazilian law does not currently permit any investigations based on the outputs of this tool (Aarvik, 2019, p. 8).

The second kind of AI applications used in anticorruption enables the automation of decision-making processes.[1] The automation AI applications instantiates is sometimes believed to reduce the discretion that unscrupulous bureaucrats may abuse to extract bribes (Santiso, 2019, p. 3). The uniform application of encoded rules may reduce officeholders' discretion under the assumption that such rules can mitigate the negative impact of corrupt personal inclinations on public institutional action (Köbis et al., 2021, p. 17). However, as we discuss later, human biases may insinuate at different levels, including in the algorithm design,

[1] We understand automation as the delegation of a cognitive or decisional function from an officeholder to some ML algorithms implemented by computing machines. For definitions of AI-driven automation in the public sector see, e.g., Loi and Spielkman (2021, p. 758).

and be perpetuated through the technology throughout the automated process and its entailments (Angwin et al., 2016; Binns, 2018; Barocas & Selbst, 2016; Tsamados et al., 2022).

AI applications are becoming a central tool in determining citizens' eligibility for such basic social services as health care, unemployment benefits, pension allocations, and child support. Such applications are widespread in the United States (AI Now Report, 2018; Calo & Citron, 2021; Eubanks, 2018) and many European countries too (Algorithm Watch Reports 2019, 2020). A further interesting example is the Indian Biometric Identity initiative, known as *Aadhar* (Arun, 2020). *Aadhar* can develop a digital identity serving as the primary portal through which citizens can get access to social services, in particular banking and food subsidies for vulnerable populations. For Payal (2019) this project stands as an anticorruption measure. The majority of India's citizens lack any form of identification (residence permits/passports); the market of fake identities thus represents a form of corruption with the negative consequence of complicating the access to welfare benefits for those most in need. *Aadhar*'s digitalization may be considered an anticorruption ally insofar as it facilitates citizens' access to welfare benefits, thus weakening the grounds for resorting to corruption to obtain those benefits.[2]

The potential contribution of AI technologies to anticorruption must reckon with the increasing concerns about their limits.[3] One concern regards users' tendency to interpret ML outputs as causative. Scientists recall that even if ML algorithms may accurately predict an event or behavior, the statistical correlations they detect between pieces of data do not prove relations of causality between them (Cowls & Schroeder, 2015; Leetaru, 2019). Moreover, ML algorithms risk to establish spurious correlations while processing large data sets, which may affect the validity of their results (Cowls & Schroeder, 2015, pp. 457–458). The possibility of errors (false positives/negatives) is always present in ML outputs, although minimized through inductive risk management techniques (Biddle & Kukla, 2017; Douglas, 2000). ML potential positive contribution should, therefore, be critically considered in view of the risks of errors in a particular circumstance or domain. But, as discussed later, the ethical challenges concerning the use of ML in the domain of anticorruption run even deeper into the normative structure of public institutional action.

ML algorithms and the limits of an exclusively legalistic approach to anticorruption

To integrate ML algorithms into anticorruption in public institutions seems justified insofar as anticorruption is conceived as targeting corrupt behavior entailing some formal rule-violation, which ML algorithms are capable of identifying or limiting. In this sense, the existing use of ML algorithms within anticorruption presupposes and implements a purely *legalistic* and *regulatory* approach.

A purely *legalistic* approach to anticorruption primarily employs retributive mechanisms for pointing out and censoring officeholders' corrupt conduct. Such mechanisms include measures for the criminal prosecution and punishment of corrupt officeholders, or their removal from office. For example, the European Anti-Fraud Office's anticorruption strategy has established a European Public Prosecutors Office (EPPO). Since the end of 2020, the EPPO has carried out cross border criminal investigations and prosecutions.[4] From such a retributive perspective, anticorruption is characteristically condemnatory. Depending on the offense, punishing measures may include the corrupt officeholders' removal from their job; their professional demotion; or imprisonment—in the case of criminal offenses such as bribery. The investigations and legal actions made possible by some such ML algorithms as *Prozorro* and *Dozorro*, as discussed in the "The integration of ML algorithms into anticorruption strategies" section, presuppose and implement such a legalistic retributive and punitive approach to anticorruption.

But anticorruption also includes preemptive *regulatory* interventions aimed at restricting the margins of officeholders' discretion to reduce the room for arbitrary uses of public power. Consider, for example, the United Nations Guide for Anti-Corruption Policies prepared by the UN Office on Drugs and Crime (2003). The Guide mentions the regulation of officeholders' discretion as one of the institutional reforms key to fighting corruption.[5] The delegation to ML algorithms of the making of certain decisions where officeholders may be called to exercise their discretionary power presupposes and implements this regulatory logic.

[2] Automation initiatives of this sort are well known in criminal justice too. One example is COMPAS, an application that has attracted much attention, partly due to allegations of algorithmic discrimination based on race (Angwin et al., 2016). COMPAS is used to assess the risk of defendants committing a general or violent crime in the future. Judges in the United States have been using COMPAS scores as an input to decide sentences (Rudin et al., 2020).

[3] For a discussion of the practical criticalities concerning the integration of ML applications into anticorruption strategies, see, e.g., Aarvik (2019) and Petheram et al. (2019).

[4] See https://ec.europa.eu/info/law/cross-border-cases/judicial-cooperation/networks-and-bodies-supporting-judicial-cooperation/european-public-prosecutors-office_en

[5] See https://www.unodc.org/pdf/crime/corruption/UN_Guide.pdf

To be sure, such measures carry some weight in anticorruption. Retributive and regulatory anticorruption strategies may tackle those cases of corruption consisting in some officeholders' abuse of their power of office in violation of some formal rule of professional conduct, and where such corrupt behavior can be identified, circumscribed, and quantified. However, many commentators have recently pointed out that such cases of corruption, while common, are not the sole, or even the predominant form of corruption in public institutions (Anechiarico & Jacobs, 1996; Ceva & Ferretti, 2021a; Mungiu-Pippidi & Heywood, 2020).

For starters, not all instances of corruption in public institutions can be remedied by preemptive regulatory actions aimed at restricting the margins of officeholders' discretion. Surely enough, professional rules of conduct in public institutions should describe as precisely as possible what kinds of officeholders' conduct are allowed or prohibited. However, an overly regulatory tendency risks stigmatizing any use of officeholders' discretion. This happens often as concerns the regulation of conflicts of interest, when, for example, rules are introduced that exclude as eligible candidates for a post in a public office those who are related (even remotely) to people currently employed in the same office (Ceva & Ferretti, 2021a, Chapter 5). This tendency can also be critiqued to express a worry for the bureaucratization of the public function. The worry is that of losing sight of the centrality of public officeholders' discretion as concerns the use of their power of office in specific circumstances. Such a discretion seems unavoidable given the complexities of institutional action, which may expose officeholders to unpredictable challenges (e.g., the management of a sanitary emergency). Moreover, a measure of discretion seems normatively desirable in order to avoid turning public officeholders into mere executors of the mechanical application of rules (we revisit this point in the "The normative challenges of integrating ML algorithms and office accountability" section).

A second concern regards the focus on corruption as the violation of some formal institutional rule. These forms of corruption are frequent as the incidence of bribery clearly illustrates. However, institutional action may as well be tainted by *ethically* relevant instances of corruption, which do not entail formal rule-breaking. This is, for example, the case of various forms of particularism in the exercise of the public function. These are cases that frequently fall within an officeholders' margins of appreciation. If a public officeholder decides to nominate a friend as a personal counsellor, they may not break any formal rule. Their conduct may nevertheless be ethically questionable if the friendship leads the officeholder to rely exclusively on their friend's advice thus cutting off the rest of the cabinet. These forms of particularism are often the result of a corrupted institutional culture. Such a culture is reveling of the officeholders' dispositions to corruption, even when they do not (immediately or ever) translate into corrupted discrete actions consisting in the violation of a rule which may be (more or less mechanically) identified, investigated, and prosecuted. An exclusively legalistic and regulatory approach to anticorruption looks insufficient on its own to make durable changes that can sustain an institutional culture of anticorruption over time.

Last but not least, the officeholders' retrospective causal or contributive responsibilities for corruption are not always clearly identifiable and assessable from a normative point of view. Consider systemic bribery or clientelism. In such cases, corruption is the product of the action of "many hands;" the very idea that anticorruption requires retributive punitive measures for the "bad apples" seems by itself insufficient (Thompson, 2017). Some problems with such an approach to anticorruption may be quite concrete; they concern, for example, the reconstruction of the chain of events factually leading to corruption. Sometimes officeholders join a corrupted institution (e.g., one tainted by familistic hiring practices) in *medias res*, when the officeholders who originally initiated the practice are no longer in service. Moreover, some officeholders may not even be aware of being involved in a corrupted practice, due of a lack of alertness or a degree of self-deception (Ceva & Bagnoli, 2021). In these circumstances, conceiving anticorruption only or mainly in terms of retributive punitive strategies risks to offer a blunt weapon or, in fact, incentivize scapegoating.

Anticorruption within an institutional ethics of office accountability

The analysis so far indicates the limits of existing legalistic (retributive, punitive) and regulatory anticorruption measures. We now propose integrating those measures within t an institutional ethics of office. This latter is necessary to incorporate into anticorruption some ethical normative guidance for public officeholders to uphold by their action the working of their institution. Our earlier discussion suggests that this integration is necessary to target corruption even when it does not involve formal rule-breaking, or when the difficulties in identifying officeholders' contribution to corruption risk to blunt anticorruption efforts. It is also important to preserve some significant margins for officeholders' discretion, while promoting their responsibility to exercise such discretion in ways coherent with upholding public institutional action (and not to their own benefit). We can now develop this claim and propose to re-appraise the contribution that ML algorithms may give to anticorruption understood in these ethical terms.

To claim that anticorruption should also include sustaining officeholders' capacity to uphold public institutional action invites the promotion of a public institutional ethics of "office accountability" (Ceva & Ferretti, 2021b).

The internal structure of public institutional action is such that, for a public institution to work, it is necessary that all officeholders partake in upholding the institution's raison d'être. An institution's raison d'être is the set of normative ideals that motivate the establishment of the institution and, consequently, its internal structure and working. It informs, *inter alia*, the terms of the mandates with which power is assigned to the various institutional roles. These roles are interrelated in the sense that institutional action requires that all officeholders exercise their power of office in keeping with their mandate because any officeholders' capacity to play their role depends on the other officeholders' doing the same (Ceva & Ferretti, 2021a, chapter 1—see also Applbaum, 1999; Emmet, 1966). So, to illustrate, for a job center to work, it is not only necessary that its statutory regulations are well designed. A working job center is a group of officeholders (including managers, civil servants, receptionists) who exercise their role-related powers (e.g., decision-making rights, confidentiality duties) in view of their mandates. Even a well-designed job center may derail if the officeholders do not uphold its action by their interrelated conduct.

The internal interrelated structure of institutional action makes public officeholders fundamentally and structurally *accountable* to one another for the uses they make of their power of office in their institutional capacity. This is the normative idea of "office accountability" (Ceva & Ferretti, 2021a). It means that any officeholders' exercise of their power of office must be sustained by the officeholders' commitment to justifying to the other officeholders the rationale of the agenda they pursue as coherent with the terms of their power mandate. An important feature of *office* accountability, that differentiates it from other forms of accountability (see, e.g., Bovens et al., 2014; Philp, 2009), is its mutuality. Office accountability is specific to the institutional context because it presupposes the normative order of right-duty relations through which public officeholders exercise the power mandated to their office. Within this context, officeholders are primarily accountable to each other for their conduct. They are *also* accountable to third parties, such as external enforcing authorities, or, in the case of democratic institutions, to the public. However, in the context of public institutional action, accountability has—also and distinctively—an inward-looking dimension.

The inward outlook premised on and required by office accountability points out an important sense in which corruption may work as an "internal enemy" (Ceva & Ferretti, 2021a) of public institutional action. From this inward-looking perspective, corruption may manifest itself as a deficit of office accountability that consists in the officeholders' failure to sustain by their conduct public institutional action. This kind of corruption occurs when public officeholders use their power of office for the pursuit of an agenda whose rationale may not be vindicated as coherent with the terms of their power mandate.

Consider, for example, a public officer in a job center entrusted with the power to contact the candidates for a job opening. One day a position as an interpreter becomes available, and the officer contacts, among the candidates, a person who happens to be an officer's acquaintance who owns a private debt to the officer. The officer contacts this person on the understanding that the income thus generated will make it possible for the newly appointed interpreter to repay their debt. This case is not an instance of bribery, nor does the officer's decision clearly violate any formal rule, especially if—as one may postulate—the appointee possesses the required skills. Nevertheless, the officer's conduct may be called corrupt as it instantiates a deficit of office accountability. Under no description, the officer's power mandate may include the recovery of a personal credit as a reason for action in an institutional capacity.

The officer may not have any issue with being accountable to the law for their conduct. The officer may well even be accountable to the public, insofar as no damage to the collectivity ensues. The officer would, however, fail the test of office accountability towards their colleagues whose capacity to perform their roles might be undercut by the officer's decision. For example, the officer's conduct may make the officer's superior incapable of justifying this, apparently random, choice in their activity report. A single officer's action may thus make the action of the job center, as a public institution, dysfunctional (Ferretti, 2018). Public institutions may thus be tainted with corruption even in the absence of any formal rule-violation. Corruption is, in this sense, also a matter of an institutional ethics of office.

To view corruption through the prism of an institutional ethics of office allows us to make two further considerations with respect to our earlier discussions. First, public officeholders' exercises of discretion are quintessential to their office accountability. The structure of institutional action exposes the working of public institutions to a degree of uncertainty with respect to the formulation of power mandates. Power mandates in public institutions are not always sharply carved, nor can they be, given the need to be adaptable to such contextual factors as, for example, shifting political priorities. Power mandates must also be responsive to such constraints on their implementation as resources availability. Moreover, power mandates are not set in stone and are susceptible to changes and revisions in view of the evolution of an institution's raison d'être over time. The capacity of public officeholders to hold each other accountable for

their action is, therefore, essential to make the officeholders engage, as an interrelated group of agents, in a self-reflective exercise about where the action of their institution is heading across institutional change and adaptation.

This understanding of public institutional action centered on office accountability requires the mobilization of officeholders to take in their own hands the responsibility for institutional action and dysfunctions, by exercising their power of office (and the discretion that comes with it) in ways capable of upholding the working of their institution. This consideration suggests also that there is always a sense in which the officeholders can be held morally responsible for institutional action and dysfunctions. Even when the officeholders' discrete causal responsibilities are hard to establish and assess from a legal point of view, as discussed in the "ML algorithms and the limits of an exclusively legalistic approach to anticorruption" section, officeholders can claim moral responsibility for such institutional dysfunctions as corruption *as an interrelated group of agents*. In our earlier example, the entire body of officeholders in the job center may thus bear interrelated responsibility for the corrupted hiring procedure. Notably, they can be held retrospectively responsible for failing to hold the hiring officer's asnwerable for the use of their power of office. But they can also be held prospectively responsible as a group for initiating corrective actions from within the institution (Ceva & Ferretti, 2021a, Chapters 4, 5).

To view public institutional action and dysfunctions through the lenses of an institutional ethics of office accountability has important implications for anticorruption too. Anticorruption becomes an integral part of the answerability practices through which public officeholders should react to institutional dysfunctions. At the normative level, this claim means that public officeholders should be prepared to engage in a critical and self-reflective exercise of analysis and assessment of their conduct and its rationale. This exercise may involve the dialogical questioning of their power mandates and the raison d'être of their institution. At the practical level, this approach reveals the need to develop tools for the officeholders' support.[6]

The question for us now is as follows: May ML algorithms be justifiably integrated into anticorruption strategies grounded in an institutional ethics of office accountability as tools to support officeholders' action?

[6] Some such tools require reconsidering institutional codes of ethical conduct (Lambsdorff, 2009), or establishing measures of mutual officeholders' oversight (e.g., internal whistleblowing; see Ceva & Bocchiola, 2018; Delmas, 2015).

The normative challenges of integrating ML algorithms and office accountability

Many ML algorithms have been accused of suffering from some degree of epistemological "opacity." The accusation concerns the lack of transparency of, and the ensuing difficulty to grasp, the logic behind specific algorithmic outputs. Opacity may take various forms and come in different degrees. We focus on three: some ML algorithms may be opaque for technical or legal reasons, or because of users' illiteracy.

Technical opacity

ML algorithms can follow a number of models such as Artificial Neural Network systems (ANNs), decision trees, Naïve Bayes, and logistic regression. Such models may differently be technically opaque. ANNs are widely spread in view of their high accuracy rates, yet they exhibit the highest degrees of opacity (Baqais et al., 2018; De Laat, 2018, pp. 536–538).

The form of opacity affecting ML algorithms such as ANNs is often indicated by the "black box" metaphor. As Pasquale (2015, p. 3) describes it, the inputs and outputs of the algorithmic system are accessible, but it is hard to inspect its inner workings to know with certitude how one becomes the other. For Burrell (2016, p. 5), the interaction between the high dimensionality of data and the technical complexity of algorithmic codes generates opacity. As discussed later, in the domain of public institutions, the ensuing partiality of knowledge may be problematic for the attribution of human responsibilities for algorithmic outputs and decision-making. There are also domains where technical opacity is not necessarily problematic. For instance, *Alpha Go ZERO*, a ML application designed to play Go, makes extensive use of neural networks and deep learning, training itself through self-play in order to identify the best moves and their winning percentages (Silver et al., 2017). However, no particular explanation of the decision-making process or justification of the selected moves seem necessary in the context of an activity whose purpose is merely winning a game (Pégny & Ibnouhsein, 2018, p. 13).

The technical opacity of some ML applications has triggered a new line of research in computer science aimed to develop Explainable Artificial Intelligence (XAI) technologies (Barredo Arrieta et al., 2020; Das & Rad, 2020; Vilone & Longo, 2020; Watson, 2022). Current methods in XAI follow two main approaches: some aim to provide post-hoc explanations of algorithmic decisions; others to ensure model transparency through the provision of core information about the algorithmic model design. Post-hoc approaches are currently dominant. They frequently employ feature attribution techniques to quantify the contribution

of particular variables to a given algorithmic output (Bhatt et al., 2020; Watson, 2022). Others offer counterfactual explanations by targeting minimal changes to the input data that would result in an opposite algorithmic outcome (Wachter et al., 2018).

XAI techniques sensibly mitigate the degree of technical opacity of some ML applications, thus countering the "black box" allegation. However, they retain some significant limitations too (Bertrand et al., 2022; Miller, 2019; Rudin, 2019; Watson, 2022). XAI techniques are unable to provide causal guarantees for algorithmic outcomes. What is more, post-hoc methods treat explanations as static deliverables of a process that terminates with a given individual algorithmic decision. Yet, in the context of public institutional action, this is an oversimplification because decision-making processes are multilayered, intertwined, and temporally extended (Miller, 2019, pp. 49–50).[7] We explore the implications of the use of XAI technologies within the framework of an ethics of office accountability in the "Algorithmic opacity and office accountability" and "How ML algorithms may weaken officeholders' engagement" sections.

Legal opacity

The most frequent sources of legal opacity are Intellectual Property (IP) instruments such as patents, copyrights, trademarks, and trade secrets (Meyers, 2019). For instance, Citron and Pascal (2014) have studied automated scoring systems for credit adjudication. They describe how IP instruments provide the legal basis to routinely deny requests for information about individual decisions. IP instruments have thus become a way to avoid audits of the underlying predictive algorithms by any actor outside the scoring entity (Citron & Pascal, 2014, p. 10). In addition, proprietary laws often cover the corpus of data used in ML algorithms training. Such laws are a source of opacity insofar as they impede a close inspection of the accuracy of the data backing decisional processes. For Citron and Pascal (2014, p. 28), while confidentiality of proprietary algorithms with respect to the public at large could be maintained, emerging legal frameworks should facilitate the disclosure of source codes to trusted neutral experts. They should also allow for audit trails to assess the inferences and correlations of specific processes. Some commentators have also noted how IP clauses have gradually changed their traditional function from protecting against a competitor's misappropriation to obstructing accountability (Katyal, 2019, p. 1246).

Illiteracy opacity

Illiteracy opacity indicates a generalized lack of technical knowledge to understand how AI decision-making systems work. This form of opacity is only partly due to the inherent technical complexity of algorithmic decision-making (Burrell, 2016; Danaher, 2016). It is by and large a circumstantial condition. Some recent surveys show an adult population "remarkably ill-informed about AI" (DeCario & Etzioni, 2021), including university students with no computer science background (Sulmont et al., 2019). Lay persons struggle to distinguish what AI applications can do in real file as against fiction, and tend to conflate AI with robotics (e.g., in such movies as *Wall-e* or *Her*, which depict robots endowed with human intelligence; see Long & Margeko, 2020). Common misconceptions include conflating human thinking with computer processing; underestimating the role of humans in AI systems design and implementation; or overestimating the power of AI to solve societal problems (Sulmont et al., 2019).

The low levels of AI literacy are not beyond remedy, of course. Moreover, in the context of public institutional action, there is a limit to what technical skills ordinary public officials need in order to appropriately interact with AI tools and integrate them into the exercise of their functions. Surely, public officials may not be expected to write and read codes. However, they also surely need higher order competencies such as understanding basic AI concepts and techniques (Long & Margeko, 2020). They should also be capable of critical evaluation, by identifying ethical concerns, as well as AI capabilities and limits in order to evaluate when it is appropriate to resort to AI, to what purpose, and in what measure. Such competencies are largely underdeveloped in the present context where AI has not yet been globally incorporated into schools' curricula or officeholders' training programs, and the main sources of information is popular media. While the number of countries offering public officeholders AI training programs is increasing,[8] the use of AI tools in the public sector is already happening, and the knowledge gap within public institutions is quite concrete.

Finally, cultural misconceptions about unrealistic levels of objectivity and accuracy of AI tools are often reinforced by communication strategies of tech corporations (Ajunwa, 2020, p. 1688; Boyd & Crawford, 2012). This phenomenon results in instances of "automation bias," which challenge not only lay citizens but also institutional structures. For example, Katyal (2019) and Wexler (2018) warn about the lack of scrutiny of courts regarding algorithmic processes in

[7] To enhance the intelligibility of high stakes decisions, some experts call for a more fundamental ex ante approach, emphasizing the need only to use ML techniques designed from the start to be fully interpretable (Rudin, 2019).

[8] Besides Finland (since 2017) and the US (2020s), in March 2022, the UK has announced a digital, data, and technology training program for senior civil servants (Day, 2022).

the context of criminal justice. Burrell (2016) and Danaher (2016) recall the urgency of granting institutional access at all levels to competent independent experts who can advise on the performance of AI tools, as well as the development of public educational efforts for AI literacy.

Algorithmic opacity and office accountability

The three forms of opacity variably characterizing ML algorithms may raise serious normative challenges for the integration of this kind of automation into anticorruption within the framework of a public institutional ethics of office accountability. The various degrees of algorithmic opacity implicated in different AI tools used in different institutional processes challenge the accountability of public institutional action. As seen in the "Anticorruption within an institutional ethics of office accountability" section, an institutional ethics of office accountability is the normative premise of a working public institution. Such a public institutional ethics demands the officeholders' engagement in a constant and vigilant communicative effort to justify to each other the rationale of their use of their power of office. Such an internal mobilization, as seen, integrates legalistic approaches to anticorruption insofar as it can engage officeholders to sustain public institutional action, and react to such institutional dysfunctions as corruption over time.

Within the framework of a public institutional ethics of office accountability, a necessary condition for anticorruption is that the various components of officeholders' deliberations about their uses of their power of office are, as much as possible, accessible and intelligible to them.[9] Insofar as, more or less significant portions of these deliberations depend on automated information processes that present some degree of at least one (possibly more) of the forms of opacity we have identified, that necessary condition for anticorruption as a matter of office accountability is variably but importantly challenged.

The history of using ML algorithms for preventing welfare frauds provides telling illustrations of how algorithmic opacity—in its diverse degrees and forms—may challenge office accountability. Such challenges emerge in their most destructive form when coupled with algorithmic malfunctions. One such malfunctions has dramatically manifested itself when the Michigan Unemployment Agency hired three private companies to develop an automated AI system, MiDAS, for detecting and adjudicating alleged frauds in unemployment benefits. Before implementing the system, public officers used to conduct interviews with claimants. The interviews aimed to explain questions and dispel doubts; officers enjoyed a large margin of discretion in asking questions and determining the presence of some fraud and how it should be remedied (e.g., by returning undue benefits, see Elyounes, 2021).

Besides reducing the costs deriving from the adjudication of benefits, MiDAS operationalizes one of the tenets of the regulatory approach to anticorruption (see the "ML algorithms and the limits of an exclusively legalistic approach to anticorruption" section). MiDASI may be seen as a tool to enhance the institutional process, by reducing the space for human error and the distortions due to officers' (ab)use of their discretion. However, between 2013 and 2015, the MiDAS algorithm incorrectly flagged over 34,000 people for fraud, causing massive loss of benefits, bankruptcies, homelessness, and even suicide (Hao, 2020). This situation resulted in a spike of court appeals and two class-action lawsuits. Investigations revealed that around 90% of the MiDAS fraud determinations were inaccurate (Calo & Citron, 2021, p. 828). The technical sources of such inaccuracies were as diverse as the misreading of scanned information; corrupt or inaccurate data mining, and errors in the "income spreading formula (Calo & Citron, 2021, p. 828). MiDAS also mistakenly flagged as fraud any discrepancies between the information provided by claimants and other federal, state, and employer records, with no margin of appreciation for unintentional errors (Felton, 2015; Wykstra, 2020). Of course, some mistakes could have occurred also by using other tools; and human errors in this department are frequent too. But the technical opacity of MiDAS made it difficult for the officers in charge to identify the mistakes and correct them. Ultimately, by progressively losing control of the institutional process, the officers' capacity to account for its malfunctions diminished too.

What is more, a number of plaintiffs proved that the agency never properly notified them of the fraud allegations in a way that could give them a reasonable chance to defend themselves (Egan & Roberts, 2021; See also 2015 official memo from Shaefer and Gray to the U.S. Department of Labor). This problem seems to apply to many automated welfare-fraud detection systems. Notifications tend to reach citizens only at the end of the process; "they don't usually give them any information about how to actually understand what happened, why a decision was made, what the evidence is against them, what the rationale was, what the criteria were, and how to fix things if they're wrong" (Wykstra,

[9] This condition is necessary, but it may be insufficient. Because institutional roles are interrelated, one officeholder may be not be able to account for their conduct for reasons that only partially concern their own deliberations. An officeholder's deliberation may also be influenced by the other officeholders' conduct. Think of cases of systemic corruption where various patterns of officeholders' interaction may initiate corrupt institutional mechanisms in which an officeholder may be implicated without even being aware of it. Moreover, forms of tainted reasoning, including self-deception, may limit officeholders' awareness of the grounds of their deliberations. Such limits do not depend on any form of opacity but derive from (implicit or explicit) cognitive failures (Ceva & Radoilska, 2018).

2020, p. 9). The opacity of the notification process is in stark contrast with the tenets of office accountability which, on the contrary, were reflected in the interview-based system originally deployed.

A measure of legal opacity made the situation worse. Interviews with leading attorneys of class-action lawsuits describe how defendants "fought hard" in court—sometimes shielding behind intellectual property reasons—to avoid sharing key information about MiDAS's innerworkings (Elyounes, 2021, pp. 495, 492). The hearings before administrative law judges often showcased some significant illiteracy opacity too, for example, facing agency staff unable to provide evidence to support MiDAS's fraud accusations, including the misrepresentation of some claimants' earnings (De la Garza, 2020, p. 4; see also the official memo from Shaefer and Gray to the U.S. Department of Labor, 2015, p. 3). The situation deteriorated to the point of having the state legislature pass a law requiring the Michigan Unemployment Agency to return to manual fraud determinations (De la Garza, 2020, p. 3).[10]

The discussion of MiDAS's shortcomings suggests that the integration of this type of ML technology into anticorruption must proceed with some precautions. Consider *Prozorro* and the Brazilian application predicting the probabilities of a civil servant being corrupt (see the "The integration of ML algorithms into anticorruption strategies" section). These are both "proprietary" applications; legal opacity may thus cover their underlying ML techniques. In the case of *Prozorro,* corruption risk indicators are not constant, but they evolve through exposure to new data inputs and correlations (Petheram et al., 2019, p. 18). Although this enhanced "plasticity" may discourage corrupt tenders from tricking the system, it adds to the technical opacity of this application as it increases the difficulties of interpreting and explaining the logic behind the process. And, of course, the challenges of illiteracy opacity are intuitively relevant in the face of the complexity of public institutional action (especially in cases of systemic corruption) and generalized lack of technical training for public officeholders.

To mitigate the opacity-related difficulties of integrating ML algorithms into the diagnosis of public institutional action, XAI techniques offer potentially appealing props for enhancing the accountability of institutional processes. For example, Loi et al. (2021, p. 262) have recently proposed that proof be given that the same algorithm be consistently applied throughout decisions ("consistency transparency"), thus enhancing the justification of decisions via XAI models based on "design publicity" (see, also, Kroll et al., 2017). More generally, post-hoc explanatory techniques can help to understand the "why" of algorithmic outputs, thus making institutional processes drawing on ML algorithms better intelligible, inter alia, to the officeholders relying on them for the exercise of their institutional functions. However, we must also acknowledge that some of the problems encountered with MiDAS would most likely still be present. XAI explanatory techniques would remain unable to detect algorithmic miscalculations or incorrect data input. Moreover, even very accurate XAI ML applications could lack normative informativeness, for example, by implementing an unfair treatment of citizens. Because algorithmic fairness is a disputed notion (Tsamados et al., 2022), public officeholders' role in assessing and answering for the compatibility of their institutional ethical standards with the fairness metric encoded in ML systems seems irreplaceable.

Additionally, court proceedings following automated fraud detection have shown some hesitancy of the human overseers to question AI-based processes (Charette, 2018). Such a hesitancy, partly due to low levels of AI literacy among officeholders, should be factored in by public institutions adopting explainable AI tools to mitigate opacity. XAI techniques may indeed exacerbate cognitive bias amongst users, including the "illusion of explanatory depth" triggered by an automated explanation in conditions of alleged process transparency (Bertrand et al., 2022).

One last remark derives from some recent studies observing how AI programmers tend to design explanatory agents in view of their own competences without necessarily weighing them for the application's intended users (Miller, 2018, p. 4). This condition makes XAI tools unable to offer interactive explanations tailored to the cognitive heterogeneity of users and their goals (Mittelstadt et al., 2019; Murdoch et al., 2019; Watson, 2022). Therefore, the officeholders' direct engagement within the framework of a public institutional ethics of office accountability seems necessary to offer normative guidance for assessing the conditions under which automated explanations are compatible with answerability practices in the context of anticorruption efforts.

Surely, the challenges we have just presented are particularly menacing for office accountability were ML applications to *replace* officeholders in resisting corruption. But public officeholders may more parsimoniously use ML applications to enhance their capacity to detect accountability deficits. For example, ML algorithms may flag those officeholders who, by the nature of their role, are most in need of their colleagues' support to ensure that their use of their power of office does not undermine institutional action. It remains nevertheless crucial that it is the officeholders' primary responsibility to investigate such deficits.

[10] Similar stories can be told of AI-driven systems for welfare-fraud detection in Australia, the U.K. and the Netherlands (Charette, 2018; De la Garza, 2020; Elyounes, 2021; Henley & Booth, 2020; Terzis, 2017).

How ML algorithms may weaken officeholders' engagement

Throughout our discussion, we have emphasized the integration of anticorruption within a generalized effort to sustain public institutional action through an institutional ethics of office accountability. Alongside legal efforts to curb corruption through regulative and retributive measures often requiring the intervention of an external authority (see the "ML algorithms and the limits of an exclusively legalistic approach to anticorruption" section), this ethical dimension of anticorruption works primarily from within an institution. It is premised on the engagement of public officeholders into a direct and continuous effort to check, sustain, and correct each other's work in the context of their interrelated institutional action. This engagement is a call for public officeholders to take on responsibility for the working of their institution and answering to each other for the dysfunctions that may nevertheless occur.

The call to action for public officeholders' engagement in anticorruption strategies can be unpacked into two elements. First, there is a commitment to enhancing the officeholders' mobilization and their capacity to ask and offer each other an account of their conduct. Second, this essential (although not exclusive) driving force of anticorruption comes from within a public institution and may not be (exclusively) outsourced. The analysis of the prospects of integrating ML algorithms into the fight against corruption should consider the contribution that this technology can make as concerns these elements too.

Regarding the first element, we see a risk of corrosion of the officeholders' engagement from decisional processes because of a growing resort to ML algorithms in public institutional action. As the automatization of critical decisions increases, officeholders are decreasingly stimulated or, in fact, capable to be vigilant, both with respect to their own conduct and to that of the other officeholders. This tendency is certainly proportional to the extent and the kind of reliance of institutional processes on ML techniques. Also, it is incrementally substantial in keeping with the level of automatization and the role of AI within it. In the case of ML applications, the tendency is also an indirect consequence of algorithmic opacity: The less officeholders are authorized to investigate (legal opacity), implicated in (technical opacity), and capable of understanding (illiteracy opacity) the grounds of the decision-making processes within their own institution, the less they can call each other to account for their conduct and answer for the failure of their institutional action.

The process of officeholders' disengagement relates to a number of background factors. These include, for example, the successful branding campaigns of tech corporations promising the accuracy and efficiency of their AI-driven applications. But think also of campaigns of development organizations fostering the modernization of bureaucratic structures through the automation of public procedures. These factors risk leading to an overestimation of ML technological innovations' positive contribution to public institutional action. Some social psychology studies suggest that such an overestimation risks making officeholders susceptible of unwarrantedly deferring to algorithmic advice, even in the face of contradictory information from other sources (automation bias, see Alon-Barkat & Busuioc, 2020, p. 24). These institutional and psychological factors contribute to sustaining a climate potentially inimical to a critical engagement not only with the practical limitations of what such innovations may or may not achieve, but also, and most importantly for us, with the ethical challenges that such innovations may raise.[11]

The increasing resort to ML risks conveying a certain presumption of distrust towards public officeholders, whose work seems in need of some "extra-human check." The propagation of such a sentiment may weaken officeholders' institutional commitment and, therefore, pave the way for further corruption. Resorting to ML instruments may even serve as a "moral buffer," motivating officeholders to relinquish their responsibility to the algorithms altogether (Busuioc, 2020, p. 832). An illustration comes from the alarm triggered by some civil lawyers in the United States about witnesses representing the state during court hearings, who are uncapable of justifying allegedly flawed decisions based on AI systems. Hao (2020) mentions such officeholders' responses as: "well, the computer did it—it's not me," while civil lawyers struggle to find effective litigation strategies and wonder with frustration: "Oh, am I going to cross-examine an algorithm?" This tendency pushes in a direction contrary to office accountability insofar as it muddles with the officeholders' claims to the authorship of institutional action. This tendency is particularly problematic from an anticorruption point of view to the extent that, as argued in "Anticorruption within an institutional ethics of office accountability" section, the fight against corruption requires engaging officeholders to answer for institutional action and dysfunctions.

The question of the authorship of decisions informed by the outputs of ML algorithms is a thorny one. Hundreds of individuals can be behind an algorithmic output, very often

[11] The 2018 European Union General Data Protection Regulation (GDPR) is one of the rare legal instruments addressing challenges arising from using AI techniques in decision-making. The GDPR requires human intervention as a safeguard on "solely" automated processing. However, it is unclear what "solely" or "human intervention" mean, thus creating a loophole by which nominal involvement of a human at any stage would mean that the decision-making process is not solely automated (Wachter et al., 2018, p. 97).

representing private institutions and their interests. There is the leadership of the tech corporation that produced the ML application; the entity commercializing the large digital datasets; the professionals assessing the quality of the data and preparing it for training the algorithm; the programmers who participate in the design of the algorithm; and, of course, the ML algorithm itself, which once developed has the potential to "learn" from new data inputs and change its decision-making rules. These manifold contributions are not only technical. They involve value choices such as deciding over tradeoffs between predictive accuracy and the interpretability of outcomes; decisions about inductive risk management[12]; and the choice of fairness metrics. Technical and value laden choices are embedded into the technology effectively institutionalizing them (Crawford, 2016; Gillespie, 2014), while the outcomes produced have the potential to affect the life prospects of millions of people. Pondering these elements in deliberation has been the traditional domain of officeholders' discretion, and a pre-condition of their taking responsibility for institutional action and dysfunctions. Insofar as the integration of ML algorithms interferes with this logic, it must be accompanied with extra caution and an awareness of the institutional transformations that may ensue.

The plurality of actors implicated in the design, development, and implementation of ML applications is also significant as concerns the second element that qualifies the integration of anticorruption within a public institutional ethics of office accountability. The compresence of many external actors and their implication in giving direction to public institutional action risks limiting the margins for officeholders' initiative to take corrective actions from the inside a public institution. A concrete form such a risk may take is that the integration of ML algorithms becomes the Trojan horse of the privatization of anticorruption efforts. Insofar as private actors hold the monopoly over automated technologies, the integration of such technologies into the working of public institutions through, inter alia, anticorruption applications may reinforce the dependence of public institutional action on third parties, thus further disengaging officeholders. Differently put, this monopoly of private actors over automated technologies opens the doors for computer programmers and CEOs of tech corporations to become de facto public institutional agents, thus triggering the privatization of many sectors of the state's action, including anticorruption.[13]

The risk of privatizing public institutional action is indirect and contingent on the design and use of ML algorithms in the public sector. The status quo may be changed, and there are some examples of ML applications conceived in a way that reduces this kind of risk. This is the case of the *Dozorro* application we encountered in the "The integration of ML algorithms into anticorruption strategies" section. The design process of *Dozorro* included a very active participation from government agencies and civil society organizations. This case suggests that the justification for integrating AI within anticorruption in public institutions is possible but conditional. It is conditional upon finding ways to pursue this innovation that do not have the effect of discharging public officeholders from their direct commitment to countering corruption by critically engaging with each other's uses of their power of office. We hope that the analysis in this section has offered an initial qualified contribution to this quest.

Conclusion

Our main aim in this artvicle has been to offer an analytical framework for understanding the challenges concerning the integration of ML algorithms into anticorruption from the perspective of a public institutional ethics of office accountability.

We started by showing how some uses of ML algorithms in anticorruption may indeed sustain the legal and regulatory efforts to constrain officeholders' conduct and identify and punish their corrupt behavior. However, given the structural complexity of public institutional action, we also argued that anticorruption strategies that exclusively employ legal and regulatory means are insufficient to sustain public institutional action in the fight against corruption over time. Alongside such initiatives, we showed how anticorruption passes also through the officeholders' direct engagement from within their institution. An important integration of anticorruption consists in offering ethical normative guidance for public officeholders' conduct. We argued that the employment of ML algorithms from this ethical perspective raises some challenges which require proceeding with a grain of salt.

Some of these challenges directly derive from the epistemological opacity that taints ML algorithms of various kinds and in different degrees. We saw how various forms of opacity may weaken public officeholders' capacity for and commitment to asking and offering each other an account of the grounds for their action, as an institutional ethics of office accountability demands. Moreover, our discussion revealed a risk that the growing resort to ML algorithms may weaken the officeholders' engagement by curbing their margins of

[12] For instance, consider trade-offs between reducing false positives (classifying an innocent as corrupt) or false negatives (classifying a corrupt individual as innocent), since oftentimes choosing to decrease the former type of errors comes with an increase in the latter type (Kearns & Roth, 2019).

[13] On the ethical risks of the privatization of state's action see Cordelli (2020).

discretion and the ensuing capacity to claim responsibility for institutional action and dysfunctions.

Our discussion has not, however, issued a death sentence for the prospects of integrating ML algorithms into anticorruption. Besides the contribution that this technology can give to legalistic and regulatory anticorruption instruments, certain forms of opacity may be mitigated insofar as they are contingent to specific practical or legal arrangements which can be changed. For example, the technical opacity of ML algorithms may be mitigated through XAI techniques, whose development and employment are fast growing yet, as seen, improvable. Moreover, IP laws responsible for legal opacity may be revised, *inter alia*, to reduce the risks of privatization of certain public functions to the advantage of those who hold the monopoly over ML technologies. Finally, certain forms of illiteracy opacity are circumstantially due to the relative novelty of ML technologies. Their ensuing shortcomings could, therefore, be remedied through a more direct involvement and education of public officeholders as concerns the design and implementation of such technologies. Automated technologies could thus serve, rather than dominate, public institutional action.

Acknowledgements Earlier versions were presented at an ECPR workshop and the GECOPOLseminar of the University of Geneva. We are grateful to the participants in those events and the journal's anonymous reviewers for the comments leading to thepresent version.

Author contributions The authors have equally contributed to the paper.

Funding Open access funding provided by University of Geneva. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable.

Informed consent Not applicable

References

Aarvik, P. (2019). Artificial Intelligence: a promising anti-corruption tool in development settings? U4 Report 2019:1. Anti-Corruption Resource Centre. Retrieved May 3, 2021, from https://www.u4.no/publications/artificial-intelligence-a-promising-anti-corruption-tool-in-development-settings

AI NOW Report. (2018). Retrieved May 3, 2021, from https://ainowinstitute.org/AI_Now_2018_Report.pdf

Ajunwa, I. (2020). The paradox of automation and anti-bias intervention. *Cardozo Law Review, 41*, 1671.

Algorithm Watch. (2019). Automating Society Report. Taking Stock of Automated Decision-Making in the EU. Retrieved May 3, 2021, from https://algorithmwatch.org/en/automating-society-2019/

Algorithm Watch. (2020). Automating Society Report. Retrieved May 3, 2021, from https://automatingsociety.algorithmwatch.org

Anechiarico, F., & Jacobs, J. B. (1996). *The pursuit of absolute integrity: How corruption control makes government ineffective*. University of Chicago Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*. Retrieved May 3, 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Applbaum, A. (1999). *Ethics for adversaries: The morality of roles in public and professional life*. Princeton University Press.

Arun, C. (2020). AI and the global south: Designing for other worlds. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI.* Oxford University Press.

Baqais, A., Baig, Z., & Grobler, M. (2018). Transparency and opacity in AI systems: An overview. In *Proceedings of the OzCHI 2018 workshop on interaction design for explainable AI*. Retrieved September 9, 2022, from https://drive.google.com/file/d/1DOSTfjKozoWi-1_b1t7aeEI2Ue03v2R2/view

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review, 104*, 671–732.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Bertrand, A., Belloum, R., Eagan, J., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (AIES'22). https://doi.org/10.1145/3514094.3534164

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 648–657).

Biddle, J. B., & Kukla, R. (2017). The geography of epistemic risk. In K. Elliott & T. Richards (Eds.), *Exploring inductive risk: Case studies of values in science* (pp. 215–237). Oxford University Press.

Binns, R. (2018). Fairness in machine learning: lessons from political philosophy. In *Proceedings of the conference on fairness, accountability, and transparency*. https://doi.org/10.48550/arXiv.1712.03586

Bovens, M., Goodin, R., & Schillemans, T. (Eds.). (2014). *The Oxford handbook of public accountability*. Oxford University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679.

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

Busuioc, M. (2020). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review, 81*(5), 825–883.

Calo, R., & Citron, D. (2021). The automated administrative state: A crisis of legitimacy. *Emory Law Journal, 70*(4), 797–845.
Ceva, E., & Bagnoli, C. (2021). Individual responsibility under systemic corruption. A coercion-based view. *Moral Philosophy and Politics*. https://doi.org/10.1515/mopp-2020-0033
Ceva, E., & Bocchiola, M. (2018). *Is whistleblowing a duty?* Polity Press.
Ceva, E., & Ferretti, M. P. (2021a). *Political corruption. The internal enemy of public institutions*. Oxford University Press.
Ceva, E., & Ferretti, M. P. (2021b). Upholding institutions in the midst of conflicts: The threat of political corruption. *Ethics & Global Politics, 14*(3), 1961379.
Ceva, E., & Radoilska, L. (2018). Responsibility for reason-giving: The case of individual tainted reasoning in systemic corruption. *Ethical Theory and Moral Practice, 21*(4), 789–809.
Charette, R. (2018). Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold, *IEEE SPECTRUM*. Retrieved September 9, 2022, from: https://spectrum.ieee.org/risk-factor/computing/software/michigans-midas-unemployment-system-algorithm-alchemy-that-cre-ated-lead-not-gold
Citron, D., & Pascal, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.
Cordelli, C. (2020). *The privatized state*. Princeton University Press.
Cowls, J., & Schroeder, R. (2015). Causation, correlation, and big data in social science research. *Policy & Internet, 7*(4), 447–472.
Crawford, K. (2016). Can an algorithm be agonistic? Scenes of contest in calculated publics. *Science, Technology and Human Values, 41*(1), 77–92.
Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy of Technology, 29*(3), 245–268.
Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint, 2006.11371.
De la Garza, A. (2020). States' automated systems are trapping citizens in bureaucratic nightmares with their lives on the line, *TIME Magazine*. Retrieved September 9, 2022, from https://time.com/5840609/algorithm-unemployment/
De Laat, P. (2018). algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology, 31*(4), 525–541.
DeCario, N., & Etzioni, O. (2021). America needs AI literacy now, *Allen Institute for AI (AI2)*. Retrieved August 26, 2022, from https://pnw.ai/article/america-needs-ai-literacy-now/72515409.
Delmas, C. (2015). The ethics of government whistleblowing. *Social Theory and Practice, 41*, 77–105.
Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science, 67*, 559–579.
Egan, P. & Roberts, A. (2021). *Judge: Companies can be sued over Michigan unemployment fraud fiasco*. Detroit Free Press. Retrieved 9 September 2022 from file:///C:/Users/didof/Zotero/storage/KKE9LX8K/7014975002.html
Elyounes, D. (2021). "Computer Says No!": The impact of automation on the discretionary power of public officers. *Vanderbilt Journal of Entertainment and Technology, 23*(3), 451–515.
Emmet, D. (1966). *Rules, roles and relations*. MacMillan.
Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
Felton, R. (2015). Criminalizing the unemployed, *Detroit Metro Times*. Retrieved September 9, 2022, from https://www.metrotimes.com/news/criminalizing-the-unemployed-2353533
Ferretti, M. P. (2018). A taxonomy of institutional corruption. *Social Philosophy and Policy, 35*(2), 242–263.
Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies essays on communication, materiality, and society*. MIT Press.
Hao, K. (2020, December 4). The coming war on the hidden algorithms that trap people in poverty. *MIT Technology Review*. Retrieved May 3, 2021, from https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/
Henley J., & Booth R. (2020). Welfare surveillance system violates human rights: Dutch court rules, *The Guardian*, Retrieved September 9, 2022, from https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules
Heywood, P. (Ed.). (2015). *The Routledge handbook of political corruption*. Routledge.
Katyal, S. (2019). The paradox of source code secrecy. *Cornell Law Review, 104*, 1183–1280.
Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
Köbis, N., Starke, C., & Rahwan, I. (2021). Artificial Intelligence as an Anti-Corruption Tool (AI-ACT) -- Potentials and Pitfalls for Top-down and Bottom-up Approaches, 1–30. Retrieved May 3, 2021, from http://arxiv.org/abs/2102.11567
Kovalchuk, A., Kenny, C., & Snyder, M. (2019). Examining the Impact of E-Procurement in Ukraine. CGD Working Paper 511. Center for Global Development. Retrieved May 3, 2021, from https://www.cgdev.org/publication/examining-impact-e-procurement-ukraine
Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., et al. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705.
Lambsdorff, J. (2009). The organisation of anticorruption: Getting the incentives right. In R. I. Rotberg (Ed.), *Corruption, global security, and world order.* Brookings Institution Press.
Leetaru, K. (2019). A reminder that machine learning is about correlations not causation. *AI and Big Data*, *Forbes*. Retrieved May 3, 2021, from https://www.forbes.com/sites/kalevleetaru/2019/01/15/a-reminder-that-machine-learning-is-about-correlations-not-causation/#5660be066161
Lima, M. S. M., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly, 37*(1), 101407.
Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology, 23*, 253–263.
Loi, M., & Spielkamp, M. (2021). Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (*AIES '21*).
Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–16).
López-Iturriaga, F. J., & Sanz, I. P. (2018). Predicting public corruption with neural networks: An analysis of Spanish provinces. *Social Indicators Research, 140*(3), 975–998.
Marzagão, T. (2017). *Machine Learning to Fight Corruption in Brazil*. (Video) Center for Effective Global Action #SmartGov.
Meyers, J. (2019). Artificial intelligence and trade secrets. *Landslide, 11*(3).
Miller, S. (2017). *Institutional corruption*. Cambridge University Press.
Miller, T. (2018). Invited talk: Explanation in Artificial Intelligence: Insights from the Social Sciences. In *Proceedings of the OzCHI 2018 workshop on interaction design for explainable AI*. Retrieved September 9, 2022, from https://drive.google.com/file/d/1DOSTfjKozoWi-1_b1t7aeEI2Ue03v2R2/view
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
Mittelstadt, B., Russel, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279–288).

Mungiu-Pippidi, A., Dadašov, R. & Fazekas, M. (2015). *Public integrity and trust in Europe*. European Research Centre for Anti-Corruption and State-Building (ERCAS), Hertie School of Governance.

Mungiu-Pippidi, A., & Heywood, P. M. (Eds.). (2020). *A research agenda for studies of corruption*. Edward Elgar Publishing.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080.

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

Payal, A. (2019). Benign dataveillance—The new kind of democracy? Examining the emerging datafied governance systems in India and China. *Communicative Configurations*, Working Paper No. 24.

Pégny, M., & Ibnouhsein, M. (2018). Quelle transparence pour les algorithmes d'apprentissage machine ? hal-01877760. Retrieved May 8, 2020 from: https://hal.inria.fr/hal-01791021

Petheram, A., Pasquarelli, W., & Stirling, R. (2019). The next generation of anti-corruption tools: Big data, open data & artificial intelligence. Research Report: Oxford Insights. Retrieved May 3, 2021, from https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5ced49ccc8302518cb27f64b/1559054797862/Research+report+2019_+The+Next+Generation+of+Anti-Corruption+Tools_++Big+Data%2C+Open+Data+%26++Artificial+Intelligence.pdf

Philp, M. (1997). Defining political corruption. *Political Studies, 45*(3), 436–462.

Philp, M. (2009). Delimiting democratic accountability. *Political Studies, 57*(1), 28–53.

Rahwan, I., et al. (2019). Machine behavior. *Nature, 568*(7753), 477–486.

Rose-Ackerman, S. (1999). *Corruption and government: Causes, consequences, and reforms*. Cambridge University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.6ed64b30

Santiso, C. (2019). Here's how technology is changing the corruption game. *World Economic Forum Blog*. Retrieved May 3, 2021, from https://www.weforum.org/agenda/2019/02/here-s-how-technology-is-changing-the-corruption-game/

Shaefer, S. & Gray, S. (2015). Michigan Unemployment Insurance Agency: Unjust Fraud and Multiple-Determinations. Official memo to the U.S. Department of Labor. Retrieved September 9, 2022, from https://waysandmeans.house.gov/sites/democrats.waysandmeans.house.gov/files/documents/Shaefer-Gray-USDOL-Memo_06-01-2015.pdf

Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. *Nature, 550*, 354–359.

Sulmont, E., Patitsas, E., & Cooperstock, J. (2019). Can you teach me to machine learn? In *Proceedings of the 50th ACM technical symposium on computer science education (SIGCSE '19)* (pp. 948–954).

Terzis, G. (2017). Austerity is an Algorithm, *Logic Magazine* (Justice: Issue 03), Retrieved September 9, 2022, from https://logicmag.io/03-austerity-is-an-algorithm/

Thompson, D. (2017). Designing responsibility: The problem of many hands in complex organizations. In J. van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in ethics.* Cambridge University Press.

Thompson, D. (2018). Theories of institutional corruption. *Annual Review of Political Science, 21*(26), 1–19.

Tsamados, A., Aggarwal, N., Cowls, J., et al. (2022). The ethics of algorithms: Key problems and solutions. *AI & Society, 37*, 215–230.

UN Office on Drugs and Crime. (2003, November). United Nations Guide for Anti-Corruption Policies. Retrieved September 9, 2022, from https://www.unodc.org/pdf/crime/corruption/UN_Guide.pdf

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. arXiv preprint, 2006.00093.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*(2), 841–887.

Warren, M. E. (2004). What does corruption mean in a democracy? *American Journal of Political Science, 48*(2), 328–343.

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese, 200*(65), 1–33.

Wexler, R. (2018). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stanford Law Review, 70*, 1343–1429.

Wykstra, S. (2020). Government's use of algorithm serves up false fraud charges, *UNDARK – Digital Magazine.* Retrieved September 9, 2022, from https://undark.org/2020/06/01/michigan-unemployment-fraud-algorithm