Minds and Machines (2024) 34:28
https://doi.org/10.1007/s11023-024-09688-8

Unfairness in AI Anti-Corruption Tools: Main Drivers and Consequences

Fernanda Odilla[1]

Received: 31 May 2023 / Accepted: 24 June 2024

Abstract

This article discusses the potential sources and consequences of unfairness in artificial intelligence (AI) predictive tools used for anti-corruption efforts. Using the examples of three AI-based anti-corruption tools from Brazil—risk estimation of corrupt behaviour in public procurement, among public officials, and of female straw candidates in electoral contests—it illustrates how unfairness can emerge at the infrastructural, individual, and institutional levels. The article draws on interviews with law enforcement officials directly involved in the development of anti-corruption tools, as well as academic and grey literature, including official reports and dissertations on the tools used as examples. Potential sources of unfairness include problematic data, statistical learning issues, the personal values and beliefs of developers and users, and the governance and practices within the organisations in which these tools are created and deployed. The findings suggest that the tools analysed were trained using inputs from past anti-corruption procedures and practices and based on common sense assumptions about corruption, which are not necessarily free from unfair disproportionality and discrimination. In designing the ACTs, the developers did not reflect on the risks of unfairness, nor did they prioritise the use of specific technological solutions to identify and mitigate this type of problem. Although the tools analysed do not make automated decisions and only support human action, their algorithms are not open to external scrutiny.

Keywords Anti-corruption · Accountability · Artificial Intelligence (AI) · Corruption · Integrity · Unfairness · Risks

✉ Fernanda Odilla
fernanda.odilla@unibo.it; feodilla@gmail.com

[1] Department of Political and Social Sciences, Università Di Bologna, Bologna, Italy

Published online: 03 July 2024

1 Introduction

MARA, Laranjômetro, and Risk Classification for Public Contracts are Brazilian artificial intelligence (AI) tools designed to fight corruption. They use machine learning to predict the risks of public officials being corrupt, female candidates being used in straw candidacy schemes, and public contracts being defrauded, respectively. These AI-based tools employ corruption risk assessment methodologies to identify preconditions, causes, and specific instances of corruption (Poltoratskaia & Fazekas, 2023). They rely on data from past events, which are not necessarily exempt from being wrongfully disproportionate and discriminatory, to make predictions and inform decision-making processes. This raises questions about whether data-driven predictive anti-corruption enforcement activities are necessarily neutral and value-free, and what potential issues may arise in terms of reinforcing inequalities.

Accountability for corruption predictive models based on risk assessment methodology has been far surpassed by its rapid advancement, and this type of technology still needs more academic scrutiny. Therefore, this article aims to explore the concept of unfairness in predictive artificial intelligence anti-corruption tools (henceforth AI-ACTs) by identifying their main drivers and reflecting on possible mitigations. It does so while assessing the three aforementioned technologies. This is particularly important for the field of corruption studies, as the critical evaluation of AI usage remains largely unexplored despite the increasing technological orientation of anti-corruption policies. With only a few exceptions (see Kossow et al., 2021; Köbis et al., 2021, 2022; Lima, 2020; Lima & Andrade, 2019), the risks associated with AI-ACTs have yet to be thoroughly examined by academics.

It is worth noting that AI applications are becoming regular tools for law enforcement and criminal justice agencies to conduct predictive analytics to fight not just corruption, but crime more broadly. These applications are believed to provide more objective and thorough decision making, as machines can process vast amounts of information at a faster pace than humans. While some analysts are excited about the possibilities of predictive automated data analysis (Santiso, 2019; Sharma, 2018) and recognise how it transformed intelligence, counterterrorism, and policing with the promise to uncover unexpected patterns (Aradau & Blanke, 2017, p. 374), civil rights and social justice groups and critical scholars highlight potential risks of perpetuating and exacerbating existing disparities (Angwin et al., 2016; Jefferson, 2018; Siegel, 2018; Edler Duarte, 2021, p.374). Critics claim that the data used to drive predictive enforcement activities are frequently limited and unfair. These predictions can be as inaccurate as those made by individuals with little or no criminal justice knowledge, further perpetuating disparities, as was the case with the notorious commercial software COMPAS that supported pretrial, parole, and sentencing decisions (Angwin et al., 2016; Dressel & Farid, 2018). Moreover, the way these tools are developed often overlooks the root causes of crime, such as structural racism, systemic disinvestment, and poverty, resulting in the over-policing of certain communities (Shapiro, 2019, p. 457).

Although issues of predictive data analytics have gained popularity in surveillance, crime, and criminal justice studies, there is still growing excitement about

the potential benefits of AI to prevent and detect corruption. While assessment of the levels of unfairness and their main sources is still scant, many governments are abandoning or being forced to ban AI-ACTs. A Dutch court invalidated a welfare fraud detection system that used personal data from multiple sources for not complying with the right to privacy under the European Convention of Human Rights, noting that there was a risk that the system would be biased against people in lower-income neighbourhoods (van Bekkum & Borgesius, 2021). In China, an AI system called Zero Trust was created to analyse extensive datasets to evaluate the job performance and personal characteristics of numerous government personnel, including information about their assets (Aarvik, 2019; Chen, 2019). However, since 2019, the Chinese system has been discontinued in numerous counties and cities, allegedly due to concerns regarding the potential occurrence of false positives and unreliable effectiveness in detecting specific corruption practices (Chen, 2019).

Therefore, there is a need for ongoing reflection on how the use of AI in anti-corruption can lead to discrimination and bias. It is important to ensure that the benefits of AI are balanced against its potential risks and that its use in anti-corruption is guided by principles of fairness, accountability, and transparency. This is so because "red flags" signalled by these AI corruption risk models may be designed based on problematic investigations and have been prompting new investigations that may result in targeted prosecutions and/or dismissals of public officeholders (Ceva & Jiménez, 2022) as well as punishment for companies and their owners.

In this article, AI-ACTs are defined as "data processing systems driven by tasks or problems designed to, with a degree of autonomy, identify, predict, summarise, and/or communicate actions related to the misuse of position, information and/or resources aimed at private gain at the expense of the collective good" (Odilla, 2023, p. 354). Corruption is broadly seen as acts that involve "the abuse of a trust, generally one involving public power, for a private benefit which often, but by no means always, comes in the form of money" (Johnston, 2005, p. 11). However, it is important to note that corruption does not solely encompass explicit breaches of formal rules and regulations, as noted by Ceva and Jiménez (2022). There are unethical practices that exploit loopholes and deficiencies within legal frameworks, while technically staying within the boundaries of the law. These practices are known as "legal corruption" (Dincer & Johnston, 2020) or "institutional corruption" (Lessig, 2013) and, as of now, have received limited scholarly attention.

Similarly, the intricate nature and risks of introducing or replicating unfairness through AI-ACTs have not been thoroughly investigated. Unlike facial recognition and algorithm-based tools used in credit scoring and sentencing decisions, AI-ACTs lack comprehensive assessment (Odilla, 2022). Aiming to address this gap, this article is structured as follows. Section 2 reviews the existing literature on emerging technologies in anti-corruption, focusing on lessons learned from security and surveillance predictive tools. Section 3 settles the theoretical basis, discussing the concept of unfairness, its main possible sources, and mitigation measures while introducing an analytical framework for assessing unfairness in AI predictive models. Section 4 outlines the research design and data collection, and explains why the three Brazilian ACTs were chosen to test the analytical framework. Next, findings

are presented and discussed, followed by the conclusion heightening that there are reasons to be concerned about the lack of transparency and scrutiny of AI predictive models in anti-corruption.

2 Unfairness of AI Predictive Systems: Establishing the Theoretical Basis

A growing literature has been suggesting that AI predictive tools, such as most of those in place to curb crime, can provide problematic outputs as they are prone to replicate existing issues such as bias and unfairness present in historical data and processes deployed to train them (Barocas & Selbst, 2016; Beigang, 2022; Kamiran et al., 2012; Singh et al., 2013; Veale & Binns, 2017). In the case of law enforcement, the topic of algorithmic fairness was brought to public attention in 2016, when ProPublica published the article entitled "Machine Bias" providing an assessment of risk predictions of COMPAS, an AI tool used to support bail and sentencing decisions in some US courts (Angwin et al., 2016). The analysis revealed that African Americans were disproportionately subjected to erroneous profiling of future criminal acts, whereas Caucasians were disproportionately rendered falsely innocent.

Since then, empirical investigations of algorithms designed to forecast the probability of criminal activity have been increasingly finding evidence of racial and geographical bias concerning individuals identified as "high-risk offenders", and racially and economically marginalised areas are mapped as "hot spots of crime" (Dressel & Farid, 2018; Edler Duarte, 2021; Jefferson, 2018; Shapiro, 2017, 2019). There is a major concern about the "dirty" data used as inputs (Richardson et al., 2019). Hence, calls for more equitable predictive digital tools in law enforcement are on the rise.

2.1 Defining Algorithmic Fairness

But what does it mean for an AI predictive tool to be "fair"? Broadly, algorithmic fairness is understood as decisions made by an algorithm—that is, the set of rules to be followed in problem-solving operations rendered into software to process input data and produce outputs (Silva & Kenny, 2018)—which should not produce unjust, discriminatory, or disparate consequences (Shin & Park, 2019). However, a consensus regarding a precise definition of algorithmic fairness has yet to be reached (Srivastava et al., 2019, p. 1). Binns (2018, p.1) explored the definition in the context of machine learning models by reflecting on whether fairness ought to be defined as guaranteeing that everyone has an equal chance of attaining some advantage, or as reducing the harm to the most disadvantaged ones. Although the author did not give a definitive answer to this, his philosophical analysis suggests that the problem lies in the way certain groups are represented in digital artefacts. Beigang (2022), in turn, noted that different moral norms are relevant for different predictions and decisions and, therefore, unfairness "might depend on factors outside the mere specification of how the algorithm moves from input data to the resulting output". To the

author, not only do predictions have the potential to exhibit systematic errors when applied to a particular group of individuals, but also decisions taken based on those analyses may lead to an allocation of resources and opportunities that contradicts the principles of distributive justice.

In their literature review on algorithmic fairness, Starke et al. (2022) identified that definitions of fairness varied, ranging from mathematical perspectives that overlap with measurement approaches to philosophical and social-science concepts of human fairness. Because these definitions are often incompatible with one another, the authors observed the necessity of harmonising concepts and measurements. The aim of this article is not to engage in an extensive discussion of the definition of fairness in AI. Therefore, to reflect on the possible drivers of the unfairness of AI-based anti-corruption tools, this study adopts a broader definition of fairness within the realm of AI (Shin & Park, 2019). However, in line with Binns (2018), Shin and Park (2019), and Starke et al. (2022), considerations of fairness should account for its calibration within the specific social context. Fairness in AI should not be assessed solely based on unequal distribution but should consider how inequality is generated and perpetuated.

Therefore, the concept of algorithmic fairness used here has an anthropocentric approach. It considers that human–machine interactions should not produce discrimination or disparate treatment of individuals or groups due to the decisions or actions made by AI systems, their developers, and/or users of these technologies. Discrimination is broadly seen here as subjecting individuals to disadvantageous conditions, including less favourable treatment, due to specific characteristics or membership of a salient social group, for example, race, age, gender, religion, and social status (Moreau, 2010; Eidelson, 2015).

2.2 Potential Sources of Unfairness

Within the realm of AI predictive systems, unfairness can emerge as a result of bias (Pagano et al., 2023) and/or noise (Kahneman et al., 2021). Bias refers to the presence of systematic favouritism or prejudice towards specific individuals or groups, which can manifest in various stages (Danks & London, 2017), such as classifier selection, feature design, training and misinterpretation of the outputs (Silva & Kenny, 2018). Noise represents random and undesired variations that arise mainly during data collection and processing, leading to inconsistent treatment of similar cases. As noted by Barocas and Selbst (2016), unfairness can also be an inherent by-product of data mining, such as the process of discovering patterns, relationships, and insights from large volumes of data. When approached without careful consideration, data mining has the potential to reproduce existing patterns of discrimination, and inherent biases from prior decisions as well as reinforce historical injustices (Barocas & Selbst, 2016, p. 674).

In addition, different data learning techniques, such as Naive Bayes, random forest, and Bayesian networks, have their own strengths and limitations. In the pursuit of optimal performance, the choice of data processing techniques can also lead to different results and involve trade-offs between various factors, such as accuracy,

efficiency, and disparity, that may impact the fairness of AI systems. On top of that, to run codes and conduct certain types of analysis, it is necessary to have a more robust infrastructure with powerful processors, expanded memory and stable online servers. Unfairness can also emerge due to the inappropriate deployment of AI technology (Kahneman et al., 2021). Even when training data is both accurate and representative, it may still capture undesirable non-statistical elements, more related to societal issues that contradict the objective of the AI tools, in other words, it may scale issues they seek to curb.

Despite rapid advancements, the current literature on fairness remains predominantly quantitative, as emphasised by Mitchell et al. (2021). Moreover, scholars often examine the main statistical and societal sources of unfairness in isolation without recognising that these sources can stem from at least three different levels: infrastructural, individual, and institutional. The primary source of unfairness may vary depending on the level, but they are not mutually exclusive; rather, they can reinforce each other, leading to adverse consequences for human judgement and decision making.

At the *infrastructural level*, primary sources of unfairness can be related to problematic data. This can occur due to errors, biases, or inaccuracies during the data gathering, storage, or pre-processing stages. Incomplete or flawed data may contain missing values, inconsistencies, or incorrect entries, which can impact the performance and reliability of AI models. Secondly, the representativeness of the data can be an issue, as the sample may not accurately reflect the diversity and characteristics of the population it is meant to serve. This can lead to biases and skewed outcomes, as the AI model may not have learned from a comprehensive and unbiased set of examples. Although we live in a datafied society, a lack of standards and limited access to certain databases, as well as to data processing devices and hardwares with insufficient capacity, may compromise the fairness of AI systems.

At the *individual level*, sources of unfairness may stem from the personal biases of those involved in the development and implementation of AI systems or their deployment in later stages to guide decision making. Consequently, individuals may consciously or unconsciously select data with undesirable properties, make decisions regarding data processing that may be considered objectionable, perpetuate past injustices, or create new ones by, for example, assigning inappropriate weights to certain factors. This risk underscores the potential for unfair outcomes, even when driven by good intentions. Moreover, users of these systems may unintentionally or intentionally be misled by the apparent accuracy and efficiency of the outputs, which can result in unfair treatment or biased decision making. In both cases, personal biases may influence choices and decisions made, encompassing aspects such as data selection, the weighing of factors, and model design in the development and use of the AI tool.

Actions are influenced by the values, beliefs, and expertise not only of individuals but also of their organisations. At the *institutional level*, various organisations, including anti-corruption agencies and police units, operate with their own sets of existing policies and norms. Institutional assumptions and choices, which are perpetuated in daily procedures and practices, can be influenced by systemic misconceptions that may perpetuate discrimination or create disadvantages. Institutional

unfairness, which can result in situations such as excessive control of individuals from lower social backgrounds or over-policing of people of colour, compromises the infrastructural level. Statistically, the data on fraud and crimes, for example, may be free from measurement errors, but from a normative standpoint, they are not. Fraud and crime rates reflect the unequal societal daily practices of law enforcement agencies.

Table 1 attempts to structure this intricate and interconnected landscape by categorising the various levels, primary sources, and primary risks of AI unfairness in predictive systems. It contributes to the existing literature by progressing towards an analytical framework with a more anthropocentric perspective that amalgamates choices, assumptions, and considerations that are not always examined together by scholars looking at unfairness and accountability in predictive algorithms.

While the analytical framework depicted in Table 1 was initially designed to facilitate the assessment of the data collected for this study, it may have broader applications in the realm of fairness in AI prediction-based decision making, as explained in Section 3. It highlights the importance of being attentive to potential risks that extend beyond automated decision making, enabling us to identify key sources and address them individually.

2.3 Mitigating Risks

The proposed analytical framework may also aid in implementing mitigation measures. Binns et al. (2017) identified one common approach to mitigating algorithmic unfairness, which involves assigning different weights to normative or ideological perspectives within classifiers that automate the enforcement of norms. Alternatively, statistical techniques are employed to ensure equitable representation of various groups, including protected ones in rankings, as proposed by Zehlike et al. (2017). Several solutions have been designed specifically to address unfairness and bias in data, such as Aequitas, AIF360, TensorFlow Responsible AI, and FairLearn (Pagano et al., 2023). While these are technical mitigation measures, they often involve human oversight. In other words, they need to involve humans "in the loop" to detect and compensate issues (Danks & London, 2017). However, as emphasised by Pagano et al. (2023), the responsibility for identifying and mitigating bias and unfairness "is entirely left to the developer", who frequently lacks adequate knowledge of the problem and needs better methodological guidelines. Furthermore, it is essential to acknowledge that not all organisations have established practices and procedures to consider certain sensitive attributes, reflect on potential sources of unfairness, and mitigate them through technological or more intuitional interventions.

In this context, the continued reliance on an over-technological approach to decision making can potentially exacerbate existing problematic practices, perpetuating their negative impact and giving rise to new issues, including those related to prejudice and discrimination. Especially in the realm of public administration, this underscores the vital importance of critically assessing instrumentation, which encompasses the selection and utilisation of tools, techniques, and methods in

Table 1 Analytical framework for evaluating unfairness in AI predictive systems

Level	Primary sources	Risks
Infrastructural	Problematic data	Data not representative, measurement errors, and missing values leading to prejudice or discrimination
	Statistical learning problem	The use of specific data processing techniques forces trade-offs related to accuracy, sensitivity, and precision, potentially leading to unbalanced analysis
	Hardware limitations	Low-quality hardware and weak data processors can lead to software bugs and compromised analysis
Individual	Personal beliefs and values	Individual decisions, influenced by conscious or unconscious prejudice and discrimination, affect AI decision making, including problematic data selection, factor weighting and/or model design. It can also lead to unfair treatment or decision making based on AI outputs
Institutional	Governance	Existing policies and norms guiding tech development and use may result in certain groups being advantaged or favoured at the expense of others
	Practices	Organisational procedures and routines that influence the development and use of AI systems may benefit some social groups while harming or devaluing others

policymaking (Lascoumes & Le Galès, 2007). While data-driven decisions may appear less subjective, our discussion has highlighted that they are far from neutral and can lead to unintended consequences regardless of their initial objectives. Furthermore, it is essential to acknowledge that the underlying rationales guiding the choice and implementation of such instruments, like other processes of implementation or evaluation, are politically driven and may aggregate other types of issues related to legal and budgetary criteria (Halpern & Le Galès, 2011), both in their formulation and their deployment. This adds a layer of complexity to the issue of unfairness.

2.4 Unfairness in AI-ACTs

When it comes to the use of AI in anti-corruption, scholars have been raising other types of issues related to opacity in their development and lack of accountability, which can render these technologies akin to black boxes (Aarvik, 2019; Ceva & Jiménez, 2022). There have also been reflections linking emerging anti-corruption technologies to the potential reinforcement of prevailing power structures (Köbis et al., 2021, 2022) and to the risk of facilitating new corruption opportunities (Adam & Fazekas, 2021). In addition, critical voices note the fact that anti-corruption technologies designed and trained to raise red flags, such as machine learning models, predominantly rely on legalistic and regulatory approaches that often focus on formal rule violations (Ceva & Jiménez, 2022). It is worth saying that, although ethically questioned, not all corrupt practices are universally regarded as illegal or involve explicit violations of formal rules. Yet formal rules cannot be flouted with impunity. Consequently, significant variations can arise depending on the specific context in which these AI tools are implemented.

Although the existing literature makes a significant contribution in terms of both theoretical and allegorical aspects, there is still scant empirical research on the topics of bias, noise, and unfairness of AI-based anti-corruption tools (AI-ACTs). Lima and Andrade (2019), Lima (2020), Starke et al. (2023), and Odilla (2023) are among the few academics who conducted empirical research on AI-ACTs. Their findings are a cause for concern.

When assessing over 30 AI-based tools developed in Brazil, Odilla found a lack of transparency and accountability in the case of governmental tools, and an overall low level of concern regarding biased code among developers. During the interviews, developers expressed their lack of concern, asserting that this technology is primarily designed to support and enhance human efforts in preventing and detecting corruption rather than making autonomous decisions.

Also, using data from Brazil, Lima (2020) and Lima and Andrade (2019) revealed that newly established companies, particularly those owned by individuals who are or were receiving a cash transfer benefit (*Bolsa Família*), are flagged by machine learning models as having a higher risk of corruption and, hence, more likely to be targeted in inspections. The authors argue that the scores attributed to these groups are unfair after conducting a thorough assessment of tools deployed by three law enforcement agencies to assess the risk of corruption in public procurement. Lima

(2020) and Lima and Andrade (2019) had access to the databases and models developed by civil servants working at the Office of the Comptroller General (CGU, *Controladoria Geral da União*), Prosecution Service of the Paraíba State (MPPB, *Ministério Público da Paraíba*), and the Court of Accounts of Paraíba State (TCE-PB, *Tribunal de Contas da Paraíba*). The databases were tested through different techniques – disparate impact remover, calibrated equalised odds, and adversarial debiased–to evaluate the accuracy, recall (sensitivity), disparity, and precision (Lima, 2020). The authors deployed the Aequitas toolkit, with metrics for fairness and biases (Saleiro et al., 2018). The findings suggest a significant disparity in the ownership of companies among individuals with low incomes, which revealed an inherent unfairness associated with the higher effectiveness of certain machine learning models.

As discussed, unfairness in AI predictive systems can arise from various sources and at different levels. In the context of corruption, for example, the presence of impunity is often observed, particularly among those in positions of power. Moreover, anti-corruption policies can be employed to justify and institutionalise political control, using them to target opponents and shield allies. Yet, it is important to note that, until now, AI-ACTs in place are more likely to be predictive technologies than generative ones. These technologies are designed to prevent corruption by tracking signals before any wrongdoing occurs, or by detecting suspicious cases that may have already happened. In both cases, data in ACTs are often sourced from various datasets, such as conviction rate or investigative procedures open for corruption-related cases. However, this poses a significant challenge due to the inherent difficulty of identifying and penalising instances of corruption. Rules, in turn, often derive from formal norms and regulations, as well as from accumulated knowledge of the most common misconducts. Therefore, rules may not include undue influences and exchanges that are not clearly illegal. Additionally, AI-ACTs can function autonomously or in collaboration with other machines and/or humans to process data, make analyses, and support human decisions (Odilla, 2023). Consciously or not, individuals have their personal values embedded in societal factors that shape power dynamics and create disparities. Public organisations, among them anti-corruption agencies, also have their formal and informal practices embedded in power dynamics even if their tasks are defined by legal frameworks. Hence, it is assumed here that AI-ACTs cannot be uncritically accepted as neutral and free of unfairness.

3 Methodology

To test these assumptions, the analytical framework for evaluating unfairness in AI predictive tools was applied, along with its core components introduced in Table 1, to the three empirical cases presented in this section. It aims to identify whether there were any instances of unfairness and, if so, to determine their respective levels and main sources. Three predictive anti-corruption tools developed in Brazil are used as case studies. These tools are designed to combat different types of corruption by assessing the risks of corruption in public contracts, identifying corrupt behaviours among civil servants, and targeting a specific form of electoral fraud.

3.1 Why Brazil?

First, the country proves to be a helpful context for the proposed analytical framework to advance the analysis of potential sources of AI unfairness and reflect on how to mitigate them. While it may not be at the forefront of global AI advancements, Brazil has made significant advances in various AI-related fields. According to a recent audit conducted by the Brazilian Federal Court of Accounts, 62% of 263 agencies within the public administration are about to implement, or have already implemented, some sort of AI system, with different levels of maturity (TCU, 2022). Although 50% of the agencies developed their tools in-house, most of the governmental agencies acknowledged a shortage of skilled personnel to develop and use AI technologies (TCU, 2022). The judiciary is leading AI development and implementation in Brazil, where emerging technologies have been introduced mainly to reduce costs and speed up internal procedures rather than deliver services or interact with citizens.

Second, Brazil has the necessary conditions for the development of anti-corruption technologies. The country has witnessed a continuous stream of corruption scandals. Because of that, it has experienced international and domestic pressure to advance accountability and anti-corruption mechanisms that have resulted in the implementation of a range of anti-corruption laws, including those that have improved data transparency and made information more accessible in digital formats (Lagunes et al., 2021; Odilla, 2023, 2024).

Third, the use of AI in anti-corruption efforts in Brazil dates to 2009, when the Revenue Service first launched its pioneering ContÁgil to automate and standardise administrative tasks conducted by its tax inspectors (Jambreiro Filho, 2019). Currently, ContÁgil supports the identification of tax fraud schemes and money laundering by reading account books and invoice sets, scanning the various data sources to which it has access, and building network graphs with people, companies, and their relationships (Odilla, 2023). Since ContÁgil, many other AI-ACTs have been developed using a wide range of types of data processing and for different purposes, mainly related to the responsibilities of the governmental agencies creating and deploying them. As noted by Odilla (2023), AI-ACT systems developed from the bottom-up are more likely to be open source and, therefore, more transparent. Conversely, governmental systems tend to be more closed due to their handling of sensitive data.

It is true that, at present, most of the ACTs in place, not only in Brazil, are predictive, in other words, dedicated to preventing corruption by assessing risks or raising red flags, rather than being generative. However, some of them incorporate generative aspects in their outputs, such as automated emails or reports summarising key findings obtained during data processing. While many of these governmental tools have the potential to undergo auditing, neither their underlying source codes nor their outputs are easily accessible to the general public. Overall, information on their use is also scant. The potential biases and unfairness associated with these initiatives and the efforts to mitigate them are still not widely understood or acknowledged.

The digitalisation of anti-corruption efforts has been faster than the capacity to critically assess all the anti-corruption technologies being developed in Brazil (Odilla, 2023).

3.2 Research Design

Three tools were selected to explore whether there are potential sources of unfairness and at which levels, according to the proposed framework, and to discuss their respective risks that could ultimately compromise the quality of anti-corruption efforts. The examples used here are AI-ACTs designed to identify different types of corruption, providing a broader scope for reflection and analysis. There is, however, a lack of open and accessible data concerning these kinds of tools, due to their sensitivity to the data they use and/or the decisions made based on their outputs. Therefore, for this article, the three cases were selected based on two main criteria: relevance and availability of open information on the tools' creation and functioning, which allows us to identify the levels of potential unfairness regarding them. Still, a significant limitation is the fact that this study only had access to information about the earlier versions of the algorithms but not to their actual codes. The three analysed tools can be summarised as follows:

(1) MARA stands for *Mapeamento de Risco de Corrupção na Administração Pública Federal* (Mapping Corruption Risk in the Federal Public Administration). It was developed in 2014–2015 by a civil servant from the Brazilian Office of the Comptroller General (*Controladoria Geral da União*, CGU[1]) as a Master's dissertation project in computer science. MARA creates an individual-level corruption score based on previous administrative proceedings, resulting in dismissals from the civil service for corrupt offences. When it was being developed, MARA's creator also tested different regression models, including Adaptive Lasso and Ridge regression, to achieve better levels of precision (81%), accuracy (83%), and sensitivity (85%) rates, according to its initial design. MARA was programmed using R and used as input the datasets available at the time by the CGU. To train the algorithm, the database of administrative convictions of civil servants was used, considering the different weights of several attributes such as salary, type of position, entry criteria to the public administration (formal exam or appointment), ownership of companies, and political affiliation. Even though not much information is available on how the CGU uses MARA in their daily activities, it became internationally known as a tool that predicts the risk of a government worker being corrupt. It is frequently cited as anecdotal evidence in texts about the use of emerging technology to combat corruption (Aarvik,

[1] The *Controladoria Geral da União* (CGU, Office of the Comptroller General) is one of the Brazilian anti-corruption agencies at the federal executive branch. It is responsible for inspecting public funds and conducting audits, imposing administrative sanctions on companies and civil servants, advancing active transparency and the right to information, establishing national networks to enhance public integrity, and encouraging involvement from civil society. CGU's employees have passed very competitive formal exams requiring candidates to hold at least a bachelor's degree, earn high salaries, and enjoy job stability and special employment rights (Odilla, 2024).

2019; Köbis et al., 2021, 2022). There are also academic and non-academic works and public presentations covering the tool's main features (Carvalho et al., 2014a, 2014b; Carvalho, 2015b, 2016; Center for Effective Global Action, 2018; Marzagão, 2017). MARA was designed to support the understaffed team of an intelligence unit at the CGU responsible for background checking, as will be detailed in the following section. The tool has always been treated as sensitive, and, hence, a very small number of workers have access to it. There is no information on whether the CGU's workers keep using it or not.

(2) The *Modelo de Classificação de Risco de Contratos Públicos* (Risk Classification for Public Contracts, RCPC) was also created by a civil servant from the CGU as part of a Master's dissertation project in 2015–2016, and it was largely inspired by MARA. The goal was to design a tool to automate the selection of public contracts to be audited, by considering the risk scores of suppliers and public contracts trained through supervised learning based on a database of companies that had public contracts and had been administratively sanctioned by the federal administration. Initially, RCPC was designed to apply logistic regression and techniques to prevent overfitting–Lasso and Ridge regression–to score risks for public suppliers and public contracts by using R's package named glmnet. Also, a multi-criteria model was used to decide which contracts to select to be audited by applying the analytic hierarchy process technique as a decision method, according to the data collected for this research. Overall, the initial design had 85.5% accuracy, 79.4% precision, and 95.9% sensibility. Its relevance has to do with the fact that it was designed to facilitate the work of CGU's auditors, and it has inspired other law enforcement agencies, such as state courts of accounts and the prosecution service, to use similar models for predicting risks in public contracts. There are also presentations and publications available detailing some of the techniques used and replicated (Carvalho et al., 2014b; Domingos et al., 2016; Grunewald & Cosac, 2016; Sales, 2016; Sun & Sales, 2018). More importantly, the tool was tested by Lima (2020) and Lima and Andrade (2019) who identified the risk of unfairness. There is no information on whether adjustments and mitigating measures were taken to improve the tool's algorithms.

(3) *Laranjômetro* was developed in 2020 by a team that included a data analyst working at the *Acredito Movement*'s joint office in the Brazilian Congress. The primary objective was to identify straw candidates running for local councillors (*vereadores*) in the 2020 municipal elections in Brazil, where candidates actively seek personal votes when they run for office, due to its open-list proportional representation system for legislative power at both local and national levels.[2] The tool was designed to identify electoral coalitions with an insufficient number of female candidates and to assess the risk of straw candidacies registered solely to meet the mandatory gender quota and access electoral funding; this is considered electoral fraud under Brazilian law and a means of embezzling campaign

[2] Brazil employs an open-list proportional representation system for both local councils and the national chamber of deputies. Under this system, parties present a group of candidates affiliated with their label but do not rank them before the election. Voters express their preferences by casting a ballot for a spe-

public funds. The feminine word *laranjas* (literally, "oranges") is jargon used to define straw people in Brazil, so Laranjômetro can be read as "Laranjas Meter". The model was designed to consider various aspects of the candidates' profiles, including educational level, occupation, age, date of affiliation, and characteristics of the municipalities where they were running for office. Its purpose was to identify individuals who had their names on the candidate list but were just pretending they were running the race to pocket funds or secure public funding for other candidates, parties, and coalitions. In terms of data processing, the project tested various classifiers. Laranjômetro employed a random forest classifier due to its 58% accuracy compared to a random classifier, which achieved only 15% accuracy (Gabinete Compartilhado, 2020). Laranjômetro's significance lies in the fact that the tool's findings regarding suspicious candidacies were shared with media outlets and the prosecution service for further analysis. Ultimately, this resulted in penalties for cases that were proven illegal. Laranjômetro, it is worth saying, is defined by its creators as a "study" not an AI-ACT.

3.3 Data Collection and Analysis

This study is part of a broader research project named BIT-ACT (Bottom-Up Initiatives and Anti-Corruption Technologies), funded by the European Research Council, which includes interviews with public authorities involved in the development and support of anti-corruption technologies. The primary objective of the interviews conducted under this research project was not to explore unfairness in AI, but rather to understand the process of developing and deploying ACTs. Participants were selected based on their experience in creating anti-corruption technologies or their involvement in the development of such tools, including those using AI. Desk research and the snowball technique, facilitated by civil servants who served as entry points, were used to identify and contact the participants. The interviews and informal conversations were conducted with law enforcement agents, many of whom are tech-savvy individuals who joined public administration due to its stability and salary benefits. They were working at the Brazilian Office of Comptroller General (CGU), Federal Court of Accounts (TCU), Federal Revenue Service (*Receita Federal*), Central Bank, and Administrative Council for Economic Defence (CADE) between February 2021 and April 2023. All interviews were conducted in Portuguese (the translation is ours). Consent was given in writing or orally.

Interviewees were asked about their personal backgrounds and motivations, their overall views on using emerging technologies to combat corruption, and their understanding of corruption, incentives, and challenges related to the creation and use of AI-powered ACTs by anti-corruption agencies. Ethical concerns regarding certain

Footnote 2 (continued)

cific candidate, and parties secure seats based on the cumulative votes garnered by all their candidates. The distribution of seats to individual candidates hinges on the number of personal votes they receive. As highlighted by Cheibub and Sin (2020, p.70), this system results in "intense competition among co-partisans and, ultimately, leads to weak electoral and legislative parties, limited public policies, regional focus, patronage, and corruption".

types of tools were also discussed. The saturation point was reached after 12 interviews and five informal conversations with civil servants. Among the participants, there were individuals directly involved in the development of the tools under analysis and others who had developed different ACTs but had some knowledge regarding the creation of the ones scrutinised in this study. Participants were fully anonymised for the data analysis, with the names of the interviewees being converted into alphanumeric characters. Apart from one written interview, the interviews were recorded, fully transcribed, and coded according to thematic analysis (Braun & Clarke, 2013) using MAXQDA Plus 2020.

During the interviews, algorithmic accountability and concerns related to the existing risks associated with these tools were also discussed. The topic of AI unfairness was salient during the interviews and, as a result, received greater attention during the initial round of data analysis. Therefore, complementary data were collected on AI-ACTs using grey literature available, including a few official reports, academic and non-academic work (public presentations and publications), and pieces of news on selected tools to explore risks of AI unfairness based on their technological infrastructure, main functionalities, and respective human interaction, following the framework introduced by Odilla (2023). Data extracted from document analysis were, therefore, combined with semi-structured interviews and informal conversations with civil servants working at anti-corruption agencies involved in the development of the AI-ACTs assessed here. A summary of the tools under analysis, including their date of creation, key features, types of data collected, and the conducted analysis, is presented in Table 2.

When conducting data analysis, it was adopted a coding reliability approach (Byrne, 2022) in which themes were developed early in the analytical process, beginning during the familiarisation with the data, which, in this case, consisted of interviews with civil servants and desk research. Summaries of what participants said were created in relation to questions regarding ethical issues, including but not limited to debates about potential risks of ACT development, existing issues with data used as inputs, and audit trails used to verify what should be considered suspicious, who holds and who should hold the tools accountable, whether their algorithms are auditable, and overall concerns about the risks that the tools may pose.

As mentioned before, the potential risks of unfairness and their main sources emerged during the interviews and their respective analyses, allowing us to, through an abductive approach, design the already presented analytical framework. The aim is, rather than to test theory (deduction) or develop theory from data (induction), to advance existing theory and facilitate the exploration of phenomena through close examination of individual cases, as Conaty (2021, p. 3) noted when discussing abduction as a methodological approach to case study research. Therefore, further data were collected regarding the three selected cases and further analysed. This process helped to identify patterns of meaning across the dataset and refine the theoretical assumptions presented earlier in this article concerning what constitutes unfairness, its main sources, and ways to mitigate it. The following section provides an analytical interpretation of the empirical data by considering the three levels of sources of unfairness applied to MARA, Laranjômetro and Risk Classification for Public Contracts (RCPC).

Table 2 Description of illustrative examples and methods for data gathering and analysis

Tool name	Creation (year, by)	Data collected	Analysis
MARA – Mapeamento de Risco de Corrupção na Administração Pública Federal (Corruption Risk Mapping in the Federal Public Administration)	2015, by a civil servant of the Office of the Comptroller General (CGU)	Academic work, including published articles and presentations (e.g., Carvalho et al., 2014a, 2014b; Carvalho, 2015a, 2015b, 2016; Marzagão, 2017)	Content was analysed to identify data inputs, data processing and data outputs, the main functionalities and potential risks
		Internal report from the CGU mentioning the tool	Content was analysed to evaluate how the anti-corruption agency applied its norms and practices in relation to the tool
		Informal conversations and interviews with two people directly involved in the development of this type of tool and with four people from the CGU who closely follow the development of ACTs in the agency	Interviews and informal conversations allowed us to explore developers' backgrounds and experiences, and how they impacted overall perceptions, including concerns and expectations of emerging technologies in anti-corruption. Also, it allowed us to understand why certain individual choices related to data selection and processing were made
Modelo de classificação de risco de contratos públicos (Risk Classification for Public Contracts—RCPC)	2016, by a civil servant of the Office of the Comptroller General (CGU)	Academic work, including articles and presentations (e.g., Carvalho et al., 2014a; Domingos et al., 2016; Lima, 2020; Lima & Andrade, 2019; Sales, 2016; Sun & Sales, 2018)	Content analysed to identify data inputs, data processing and data outputs, the main functionalities and potential risks
		One interviewee directly involved in its development and five people who are familiar with the tool or have developed similar ones	Interviews allowed us to explore civil servants' experiences, and how they impacted overall perceptions, including concerns and expectations of emerging technologies in anti-corruption

Table 2 (continued)

Tool name	Creation (year, by)	Data collected	Analysis
Laranjômetro (straw female candidate meter)	2020, by the joint office of the *Acredito Movement* in the Brazilian Congress	Reports detailing the tool (Xavier, n.d.; Gabinete Compartilhado, 2020) One informal conversation Pieces of news	Content coded according to thematic analysis to identify data inputs, data processing and data outputs, the main functionalities and risks. These reports highlighted the existence of bias and actions to mitigate it

4 Exploring Levels and Main Sources of Unfairness in AI-ACTs

Understanding how AI-ACTs are created by identifying the main motivations behind their development, and what their main features and functionalities are, is a crucial undertaking in understanding the complex dynamics and implications of unfairness in AI in anti-corruption. By delving into different levels, e.g., individual, institutional, and infrastructural, we can shed light on the multifaceted nature of unfairness and possibly identify its origins. The remainder of this section provides a nuanced description of the types of anti-corruption tools used as examples in this article. The goal is to identify and, if possible, assess the three levels of unfairness and their possible main sources and implications according to the framework introduced in SubSection 2.2.

4.1 Individual Level

Personal premises and motivations play a significant role in influencing the choices and decisions made throughout the development and utilisation of AI tools. These premises encompass various aspects, including data selection, weighing factors, and model design. Developers and designers of AI systems bring their own beliefs and values into the decision-making process, which can shape the outcomes and behaviour of the tools they are creating.

The case of MARA is telling in this regard. The application was created by a civil servant who saw the opportunity to combine professional responsibilities with his goal of pursuing a post-graduate degree in computer sciences. At that time, the civil servant was working at the CGU in an understaffed department, despite its numerous duties. The department, among other things, was responsible for background checks of individuals before their nominations for high-level positions and for supporting investigations of possible wrongdoings committed by federal public servants within the executive. According to Interviewee CS0X_INT009, the idea of MARA came when the civil servant was writing a preliminary proposal for acceptance on a Master's programme. It was approved by the CGU because, despite being an individual academic initiative, it had the potential to produce something useful to the anti-corruption agency as an outcome.

The interviews suggest that unfairness was not a topic under discussion when the tool was conceived. For example, it was a deliberate choice of its creator to leave out more personal information such as gender, education level, and age, preferring to focus on more professional aspects such as time working as a civil servant, salary, and whether the civil servant was occupying a position of trust or not. However, there was also the decision to include party affiliation as part of the algorithm. The use of party affiliation data to measure the corruption risk of public officials later became an academic paper (Carvalho et al., 2014a). Although MARA was an individual initiative, Interviewee CS0X_INT009 noted that many decisions involving its creation were not taken only by the civil servant who developed the application, in an attempt to mitigate potential issues. However, most of those who participated have a background in computer science and engineering. There were weekly

meetings to discuss decisions taken with CGU workers, and some of the findings were tested with professors at the University of Brasilia to validate them. In addition, it was decided to use a lower recall, which measures the completeness of positive predictions, but a high precision, which in turn signals the proportion of positive identifications that are actually correct. In addition, there were efforts to apply different statistical techniques and validate findings.

However, the tool is not free from internal critiques. Interviewee CS0X_INT009, for instance, recognised as an issue the fact it was used to label as "corrupt" the small database of civil servants punished administratively with dismissal for corruption and to label as "noncorrupt" all the others, including people who received other types of punishments, such as suspension and fines. MARA's initial model had a great focus on specific governmental units. Some of them are notorious for having a high rate of punishment, but also for having street-level bureaucrats highly exposed to corruption, such as the Social Benefit Service (*Instituto Nacional do Seguro Social*, INSS), as noted by Odilla (2020) when analysing the uneven distribution of administrative sanctions for corruption among federal agencies in Brazil.

Still, MARA was the first predictive model created at the CGU, and its rationale inspired many other models, such as supplier risk score and contract risk score. As one of the interviewees noted, it became "case-based learning":

> "So, it [MARA] became a kind of learning case, on how to use algorithms, [statistical] tools, and everything else. It served to help many other projects that are a copy of MARA but applied in other contexts that are easier to apply [as a predictive tool]." (Interviewee CS0X_INT007)

The Risk Classification for Public Contracts (RCPC, henceforth) is one of the ACTs created after MARA, following very similar principles. The initial version of this tool, designed to assist auditors in identifying high-risk cases within public contracts and calls, was trained using historical data from companies that had previously engaged in irregular activities when contracted by the public administration and had been sanctioned. The creator chose to use data from sanctions imposed in 2015 and 2016, along with contracts and bids from 2011 to 2016. The relatively short time frame could cause issues related to, for example, the representativeness of the sample. In fact, the risk qualification criteria for companies resulted in a total of 723 companies previously classified as high risk and 41,287 companies previously classified as low risk. To balance the dataset, the developer, who conducted an extensive analysis to minimise statistical learning problems, decided to apply a process known as undersampling, randomly selecting an identical quantity from the low-risk companies to match the number of high-risk ones.

Like many risk-scoring tools developed for curbing corruption in public procurement, the algorithm of the RCPC tool also considers several dimensions, such as the company's operational capacity (e.g., number of employees, partners' occupations, and whether partners receive or received social benefits), connections with politicians and campaign financing, and competition capacity (e.g., average of bids and success rate in securing public contracts). One noteworthy aspect of these tools is the consideration of specific criteria for identifying a supplier company as high risk: involvement in addictive contracts, seen as a form of circumventing new bids.

Despite the efforts to diversify the databases and incorporate important factors associated with high risk, the statistical choice was a logistic regression in which the main criterion to be classified as high risk was having past sanctions. This individual decision may have left out many other possible interventions or options, including cases of companies that engaged in wrongdoings but were not caught or punished.

Interviewee CS0X_INT007 has been studying and helping to develop predictive AI-anti-corruption tools in public procurement. Clearly inspired by money lending scores, the participant explained what motivates the creation of these types of tools designed to identify the risk of corruption among suppliers as well as in public contracts:

> "So, it's like what banks do when you apply for a loan. They ask you if you're married or single, your age, if you have children or not, if you have a job, and what your salary is. Based on all this information, the bank gets an idea of the risk you pose when taking out that loan, whether you're more likely to default or not. Then, based on this risk assessment, they give you an interest rate. So, this work we did, the risk map of suppliers, followed a similar approach, trying to quantify and predict the risk of a company encountering problems with the government based on variables related to the company." (Interviewee CS0X_INT007)

Interviewees acknowledged that discussions on AI bias and unfairness occasionally arose, but mainly in informal settings. Unfortunately, governmental agencies lack specific guidelines or frameworks to address these issues when developing anti-corruption solutions. One important exception is the project Laranjômetro, which openly pointed out the unfairness of its algorithm. On his webpage, the developer responsible for the AI-ACT explained what happened with Laranjômetro:

> "We created a machine learning model called random forest to classify female candidacies as 'likely straw candidates' and 'likely legitimate candidates'. In this process, we identified several characteristics that, when combined, are more common among straw candidates (…) We estimated that there were at least 5,000 straw candidates nationwide in Brazil for the 2020 elections, although our model only identifies a small portion of these candidates. Additionally, we identified approximately 700 coalitions that did not meet the minimum required number of eligible female candidates, according to the law. We also found that the model exhibits algorithmic bias: When selecting potential straw candidates, it more frequently selects younger, less educated candidates from the northeast region and from lower socioeconomic backgrounds (even when compared to the actual profile of straw candidates). Furthermore, it also selects Black candidates more frequently than their actual representation among straw candidates, even though the model does not explicitly use race as a feature." (Translation ours, Henrique Xavier, n.d. Retrieved May 30, 2023, from https://henriquexavier.net/laranjometro.html.)

To mitigate this issue, Laranjômetro's creators decided to apply weighted resampling to the candidates. They also tried to validate the tool with the help of journalists.

Reporters received a list of 50 female candidates and contacted 20 of them. Three of these cases were confirmed to be straw candidates (Turtelli, 2020) but there is no information on the others. In the case of public procurements, Lima (2020) highlighted that, although there might have been a trade-off between reducing unfairness in sensitive environments and losing the efficiency of the models, this shows that there are some toolkits that would help developers to mitigate the risks. Interviews with the developers, however, indicated a low level of concern, as already pointed out by Odilla (2023), as they argued that these predictive tools do not take autonomous decisions but only support the work of inspectors and their audits.

4.2 Institutional Level

The MARA individual-level corruption risk score was created as an attempt to alleviate an organisational problem—limited staff—and to automate decisions in order to reduce subjectivity in certain procedures within the CGU. MARA was developed in the context of increasing institutional support within the CGU to invest in what was then called data mining to make better use of available digital resources, relying on machines to be faster and more accurate than humans in preventing corruption. Not by chance two interviewees—who were part of the team of data scientists set up by the agency at the time—compared the MARA tool to the film (inspired by the novel) *Minority Report*, in which a police programme called "Precrime" has been in operation for six years, using a prototype system to foresee future homicides. Law enforcement officers analyse the visions of the "precogs" with psychic skills to identify the location of the crime and apprehend the potential perpetrator before the act is committed. Once identified, individuals who are predicted to become killers are placed in a state of induced coma using electrical stimulation.

MARA was designed according to the same basic criteria that an auditor would consider when conducting the analysis manually, based on the prevailing understanding in the CGU of the most common predictors of a potentially corrupt official. This does not mean that it was wrong, nor that it was free of institutional power dynamics. To illustrate this, Interviewee CS0X_INT009, involved in the creation of MARA, shared an anecdote: During a meeting to discuss auditing priorities and criteria, someone advocated for assigning greater weight to a specific variable. When asked for the rationale behind this decision, the response was simply, "Because my boss said so", and this was met with no objections from others.

Although the anecdote suggests that even the non-AI-based procedures at the CGU may have their own bias and unfairness issues, the exact use of MARA, however, remains unclear. Not every interviewee from the CGU, including those involved in developing AI-ACTs, claimed to have knowledge of the tool. Some interviewees mentioned hearing about it, while others expressed concerns that it may be operating covertly. Some individuals stated that the tool has not been widely adopted and that it is not used daily due to legal restrictions.

In its initial version, MARA was used, according to people directly involved in its deployment, in a very careful way. In addition, only a small number of people had access to it. MARA's use was detailed in one interview:

> "It's not because someone scored high in the model that a disciplinary administrative proceeding (PAD, *Procedimento Administrativo Disciplinar*) would be initiated. (…) The model was only used for the initial selection process. It depends on the context. For instance, during an audit or a special operation it would check the possibility of a person being connected to the organisation being investigated. [If yes], then, here, we would request the declaration of assets (...) Only one department had access to MARA—actually a few people from the directorate. [The information produced by MARA] was not included in any [written] reports or documents. Precisely because we were aware of the problems it could cause in terms of people not understanding the purpose of the tool. It was used internally. (…) We would discuss the risks identified only with the person who requested a report internally at the CGU, but we wouldn't include it in the report." (Interviewee CS0X_INT009).

Interviewee CS0X_INT008, however, shed light on the risks of operating this type of risk score and profiling tool surreptitiously:

> "I tend to think that if there was any decision to stop using the tool, it was not due to ethical reasons. But that's also a problem, right? Sometimes the areas that are dealing with these tools are not used to a culture of openness, participation, and transparency." (Interviewee CS0X_INT008)

Additionally, it is noteworthy that a significant level of trust was placed in the procedures and internal norms that guided non-AI-related activities, which were subsequently incorporated into the algorithms. Undoubtedly, this increases the likelihood of inheriting various institutional issues, including the potential risk of unfairness. One such example is the use of a database curated by the CGU that relies on administrative punishment after peer investigation as a proxy for determining whether a civil servant is corrupt. The interviews suggest that MARA does not address, for example, the significant disparity in the enforcement of sanctions across different agencies, both in terms of corruption and non-corruption penalties.

In fact, a study conducted by Odilla (2020) on this same administrative punishment database showed a considerable variation in corruption control when the distribution of sanctions was combined with an analysis of interviews with civil servants responsible for investigating their colleagues or coordinating administrative procedures. The author concluded that civil servants, who in Brazil are also responsible for investigating their peers at the administrative level, openly express discomfort in their role as corruption fighters, and many of them often exhibit self-protective behaviour. The punishment database, as argued by Odilla, reflects what she refers to as "convenient accountability"–a form of institutional abdication combined with a reluctance for peer monitoring. The outcomes of this approach can be described as merely satisfactory for integrity agents, rather than effectively addressing the issue of corruption in the Brazilian federal executive power.

For RCPC, it could also be observed in the documentation available that, as expected, certain auditing practices and procedures were incorporated in algorithmic modelling. For example, anti-corruption agencies commonly employ certain criteria to identify façade companies when scoring the risk of corruption in public procurement. They include high-risk factors such as the company's small number of employees, the occupation and social status of its owners (which may not be related to, or could be seen as incompatible with the company's operations), the location of its premises (such as being situated in a disadvantaged area or in unmarked buildings), and date of creation (newly created companies may be formed just to participate in a certain bid, for example). These indicators are based on previous experiences and findings and are believed to help inspectors to detect potential cases of companies that may be fronts for corrupt activities or engage in fraudulent practices in the procurement process. When used as an indicator of high risk, however, it can not only perpetuate the targeting of a particular type of irregularity but also create unfairness towards individuals and businesses in underprivileged areas who may face a higher likelihood of being inspected.

In the case of Laranjômetro, although developed by people working for elected representatives in Congress, its creators could not take further action with their findings and, therefore, they not only asked the help of journalists to validate the findings but also shared their analysis of straw candidates with law enforcement agents (Xavier, n.d.). Laranjômetro served as a basis for the inspection carried out by the Federal Public Prosecutor's Office in 2020. Laranjômetro has not been, at least until now, converted into any official mechanism used by the electoral court or the prosecution service to investigate candidates and parties. In this case, we can see how governance limits the use of the ACT.

4.3 Infrastructural Level

Undeniably, data use and processing can be pointed out as substantial sources of potential unfairness in AI-ACT, as happens in police data (Richardson et al., 2019). This is primarily due to the elusive nature of corruption, making it challenging to identify, quantify, and categorise certain practices or combinations of attributes as more likely to be associated with corruption. As previously mentioned, risk-scoring systems for corrupt civil servants and companies in public procurement often rely on training their data using punishment records, assuming a standardised profile that will persist over time. This premise not only overlooks the dynamic nature of corrupt practices but becomes problematic if it is done without assessing variations of both sanctions and corrupt behaviour, and without employing techniques to mitigate potential issues related to the manner in which investigations are conducted and the profile of those sanctioned. This approach is also inherently problematic as it fails to account for individuals who have engaged in malpractices but have not been punished.

In the CGU, the risk score tool for public procurement RCPC focuses on suppliers and was developed in six phases (Grunewald & Cosac, 2016; Sales, 2016). Initially, the model employed a pilot approach, using 1,446 companies divided equally

into high-risk and low-risk categories. This training process involved 46 predictor variables and a dependent variable linked to previous legal sanctions outlined in federal legislation, such as temporary bid suspension, prohibition from signing public contracts, and disreputable declarations. Then, a database was created, with separate datasets for testing and learning purposes. Using a stepwise algorithm, 29 variables were selected and sequentially entered into the model. Logistic regression was employed to create the final model using the training dataset. Considering the levels of accuracy, sensitivity, and precision of the models, the results indicated that higher risk was associated with the size of financial donations during elections,[3] companies registered for a wide range of activities, a smaller number of employees, lower partner salaries, and recently established companies. According to the findings of Lima (2020) and Lima and Andrade (2019), when employing methods to detect and address unfairness, it became evident that newly established businesses, especially those owned by individuals receiving or having received social benefits, were categorised as higher-risk entities. The underlying assumption is that individuals from lower-income backgrounds or recently formed businesses are more prone to being involved in straw ownership and engaging in shell company practices, respectively.

MARA also followed a similar development process, utilising the CRISP-DM (cross-industry standard process for data mining) framework. This approach, created in the 1990s, involved several stages, including understanding the business context, data understanding, data preparation, modelling, evaluation, and deployment of the tool. Initially, four risk dimensions were defined: punishment for corruption (e.g., based on administrative dismissals rather than court decisions, on electoral sanctions, etc.), professional attributes (such as occupying a position of trust, receiving an additional salary, total salary, number of positions held, number of governmental agencies worked for, tenure in each position, active career civil servant, retired status, political appointee, etc.), political factors (affiliation with a political party, duration of party affiliation, reasons for cancellation, party name, having run for office, having received campaign donations or donated to parties and candidates, etc.), and business-related indicators (ownership of a company, contracts with the public administration, number, and type of work/services/goods delivered, eventual sanctions, etc.). Different databases were joined, and data were cleaned and standardised to create thousands of attributes tested regarding their correlation and variance to then run sampling techniques and regressions (Ridge, Lasso, and Adaptive Lasso). The model was left with 32 attributes and a constant to be trained using Class-Attribute Interdependence Maximization (CAIM), Minimum Description Length Principle (MDLP), and ModChi2. Robustness tests were run.

In the case of MARA, the information available on its initial version indicates the features are more likely to be attributed to individuals with a higher risk of corruption. They include—but are not limited to—holding or having held specific

[3] In its first version, the initial decision was to add all donations made by each company in 2010, 2012, and 2014 and evaluate the odds ratio related to contract risk. In practical terms, this approach did not create different categories based on the size of the donation but made an overall estimative. The analysis indicated that for every R$ 100,000 in donations (a common amount for companies), suppliers increased their risk of contractual issues by 1.42%.

positions of trust (however, not the top ones), being previously investigated or targeted by administrative procedures at the CGU's Internal Affairs Unit, being or having been affiliated with a political party, and being part of business partnerships. It is not known whether all the attributes listed in academic work, for example, were kept in the model. One of the interviewees remembered that "having been previously investigated" was not considered. In any case, the model uses its outputs as inputs to be trained again. The unfairness in the case of MARA needs to be better investigated, but at first glance, what it indicates is the risk of being punished for corruption and not necessarily for being corrupt. Interviewee CS0X_INT009 said that nowadays there are techniques, such as PU learning, also known as positive unlabelled learning, which considers that you have a positive label, and the rest is unlabelled. According to him, this would compensate for the fact that the model has only information on the characteristics of who was labelled as corrupt, but not the others about whom "we do not know whether they are not corrupt or simply have not been caught yet".

In the case of Laranjômetro, an open report details both data used as inputs and data processing techniques (Xavier, n.d.; Gabinete Compartilhado, 2020). The developer decided to use data from four local elections since 2004 as an attempt to predict female straw candidacies in 2020. The model assumes that: (1) all female candidacies were regular before 2009, when the minimum quota of 30% for each gender was introduced, and (2) the fraction of regular candidacies that receive up to one vote remains stable with the introduction of the quota. Based on these assumptions, they estimated that, out of all female candidacies with up to one vote after 2009, 94% are straw candidates. To assess the candidates with one vote before and after the new rule, the model considers 35 features by using data from the electoral court (personal data of the candidate, declaration of assets, number of seats in each local council, the profile of voters in each municipality, votes the candidate had received in previous elections), from the database of individuals affiliated with political parties in Brazil, the human development index, and political alignment of parties and government.

According to the report, to build a classification model (supervised learning) for female candidacies for the position of councillor as "*laranjas* (straws)" and "regulars", the Laranjômetro only used data from the municipal elections of 2012 and 2016. Together, they totalled 273,669 candidacies, with 48,253 of them receiving up to one vote (corresponding to 17.6% of the total). The one vote was a deliberate choice, a criterion used to classify straws by considering that only that person would vote for herself. The data were divided into three disjointed subsets: training sample, validation sample, and test sample. To prevent the model from adopting strategies and characteristics that are not stable over time (i.e., that are not reproducible from one election to another), it was used the 2012 data as the training sample and subsets of the 2016 data as the validation and test samples. The chosen classification method was the one that best predicted straw candidacies in 2016, based on examples from 2012: a random forest model. Its precision (fraction of classified straw candidacies that were actually "*laranjas*") was 63% in the test sample. Then, based on the most common straw characteristics identified, an analysis using the 2020 database of candidates was conducted to identify outliers and potentially new types of candidates

that had not been identified before. Finally, the impact of the new legislation introduced in 2018 to transfer a minimum of 30% of public funds to female candidates was analysed. The distribution needs to be proportional to the number of candidates, but the outputs suggested a level of unfairness:

> "Compared to candidates who typically receive more votes, the profile of '*laranja*' candidacies tends to be less White, less educated, younger, and residing in poorer municipalities. At the same time, machine learning models tend to generalise the label (in this case, *laranja*) to similar examples (in this case, candidacies with the mentioned profile). This introduces bias in the model, which frequently selects candidates with this mentioned profile excessively." (Gabinete Compartilhado, 2020).

The creators of Laranjômetro recognised that the AI-ACT could create issues for individuals who are typically more vulnerable and are not straw candidates. Still, the predictions were tested with the 2020 ballot results and it was observed that 64% of the candidates selected as straws by the model received a maximum of 10 votes, while this fraction is 20% for the non-selected candidates." The people responsible for the tool, however, admitted that a detailed investigation would be necessary to qualitatively evaluate whether the straw candidates identified by the model were straw candidates or someone who received a small number of votes.

5 Discussion

Opacity was observed as a prominent characteristic of all three tools analysed: their codes are not available, and in the cases of MARA and the Risk Classification for Public Contracts, there is no information regarding updates or specific details on how these tools are utilised by the Brazilian anti-corruption agency within the federal executive. Even Lima (2020) and Lima and Andrade (2019), who were granted access to databases to evaluate and identify unfairness in the risk-scoring tools for monitoring public contracts, do not disclose the tools' inputs and outputs. Lima and Andrade (2019) stated that the tools evaluated "do not provide a detailed description of all the characteristics used by regulatory bodies to monitor companies and contracts. These bodies consider the confidentiality of defining these characteristics as strategic". Although the sensitivity of AI-ACTs should be a topic of concern, their overall lack of transparency or any other mechanism to allow external scrutiny compromises any in-depth assessment. It is essential to design AI systems that can be contested, allowing for human intervention at various stages of their lifecycle, to curb potential harm caused by automated decision-making (Alfrink et al., 2022).

Based on the data collected, the findings suggest strong indications of potential unfairness at all three levels. However, only in the case of Laranjômetro were issues related to the tool's outputs exposed by its developers. This is not to suggest that the creators of the two CGU tools, MARA and RCPC, did not attempt to mitigate risks when developing the first version of the tool. Their efforts were more focused on utilising statistical techniques to reduce systematic discrepancies between the sample used to train a predictive model and, in the case of MARA, consulting experts to

select parameters–but in both cases it was not observed any specific intervention to mitigate unfairness.

Still, at the infrastructural level, all three tools encountered statistical learning problems and did not incorporate, to the best of my knowledge, solutions specifically designed to mitigate unfairness from the outset. In the cases of MARA and RCPC, it was evident that problematic data were used, as both ACTs were trained with a database of sanctions that may not represent the entire population of corrupt civil servants and companies. They treated all other cases as noncorrupt, without acknowledging that they might not have been detected and punished yet. As mentioned, Laranjômetro also utilised past election data if no one had been a straw candidate before. However, at least Laranjômetro recognised the unbalanced outputs and attempted to mitigate them by seeking external validation from journalists and the prosecution service.

At the individual level, it was evident that many choices were influenced by the personal beliefs and values of the creators, shaped by their own perceptions of corruption and their experiences, particularly in the case of the civil servants who developed and later used MARA and RCPC. This brings us to the institutional level, where formal and informal organisational practices and routines, often based on past experiences and regulations, allowed them to decide to use indicators such as political connections of both civil servants and companies as risk indicators, without any formal guidelines to mitigate unfairness. Table 3 summarises the findings after applying the analytical framework introduced before.

Table 3 shows how levels and sources of unfairness may be interconnected, reinforcing each other and contributing to persistent inequalities and social injustices. Despite the limited data available, findings are robust enough to assert that AI-ACTs cannot be uncritically accepted as neutral and free from unfairness. Engaging in open reflection on the methods of data collection, processing, and utilisation is crucial to establishing the legitimacy of both the datasets themselves and the law enforcement practices responsible for their development and use in decision-making processes. Furthermore, it is essential to encourage open discussions of these risks among developers, as exemplified in the case of biases against young Black women, already identified by the creators of the Laranjômetro project (see Xavier, n.d.; Gabinete Compartilhado, 2020). Users should also be encouraged to evaluate, whenever possible, potential sources of unfairness to avoid the reinforcement or creation of new types of prejudice and discrimination. Although the focus is on ACTs, the insights provided by this article can be expanded upon and help advance the discussion of the risks of predictive AI systems as they are widely adopted by the private sector and rapidly advancing into government domains (Mitchell et al., 2021).

6 Conclusion

This article discusses the potential sources of unfairness in AI predictive tools applied to anti-corruption efforts. It uses three examples from Brazil to illustrate how unfairness can manifest at the infrastructural, individual, and institutional levels, with potential sources related to problematic data, statistical learning problems,

Table 3 Main sources of unfairness and potential risks observed in MARA, Risk Classification for Public Contracts (RCPC), and *Laranjômetro*

Level	Primary sources	Cases	Examples of identified risks
Infrastructural	Problematic data	MARA	–Data on affiliation from the Electoral Court was not updated frequently. It also required cleaning to address inconsistencies that may have led to missing data –Data on sanctions against companies had issues related to the timeframe of the sanction –Data on electoral donations had to be cleaned to address inconsistencies in the size of funding, which may have led to missing data –Use of subsamples of civil servants classified as noncorrupt that could have engaged in corruption but not caught
		RCPC	–Potential sample issue, as it considered as "high risk" the features of companies sanctioned only in 2015 and 2016, which represents a relatively short time frame –Considered the number of owners of companies who are civil servants, but there might be missing data, especially in cases of beneficial ownership
		Laranjômetro	–Considered personal assets declared by candidates that may not be accurate, giving greater weight to those who presented lower values and fewer items –Considered recent political affiliations without evaluating previous affiliations
	Statistical learning problem	MARA	–Model trained with a small database of punished civil servants (fired for corruption) without accounting for issues with administrative procedures, excluded variables that may have a strong weight, despite the efforts to address statistical issues
		RCPC	–Model trained with a relatively small database of sanctioned companies and a subsample of non-sanctioned ones, without accounting for those that were not caught and punished –Despite efforts to address statistical issues, parameters related to recently established companies and those owned by former beneficiaries of social benefits were not weighed or balanced
		Laranjômetro	–Considered all female candidacies as regular before 2009, and the proportion of regular candidacies that received up to one vote should remain stable with the introduction of the quota. It identified issues related to fairness
	Hardware limitations	NA	NA

Table 3 (continued)

Level	Primary sources	Cases	Examples of identified risks
Individual	Personal beliefs and values	MARA	–Individual decisions to include political affiliation and exclude gender, age and educational background as relevant features for corruption
		RCPC	–Individual decisions to include as parameters of suspicious behaviour beneficiaries included recently opened companies
		Laranjômetro	–Individual decisions to consider as suspicious shortened ballot names, no assets statements, and candidates who presented themselves as students and housewives
*Institutiona*l	Governance	MARA and RCPC	–Absence of internal norms requesting the use of systems designed to mitigate unfairness and bias –MARA classified as "corruption" specific legal justifications that lead to civil servants being terminated on the basis of the interpretation of internal reports from the CGU –RCPC considered CGU's tools that reflag suspicious cases of public bids and their participants, risking replicating already existing issues
	Practices	MARA and RCPC	–Informal discussions revolved around the utilisation of certain parameters, indicating institutional understandings regarding what constitutes suspicious behaviour

personal values and beliefs of those developing and using these tools, as well as governance and practices within the organisations where they are created and/or deployed. The findings corroborate the existing literature, which has already highlighted the opacity of AI-ACTs, compromising their auditability and increasing the risks of bias and noise (Ceva & Jiménez, 2022; Odilla, 2023). This is particularly the case for tools deployed by governments that train their models based on sanctions, such as in the case of MARA, which assesses the risk of civil servants being corrupt, and Risk Classification for Public Contracts, RCPC.

Furthermore, from interviews and descriptions of the initial tool versions, it becomes evident that despite attempts to address statistical issues using established techniques, there has been not only limited access to AI-ACTs codes but also limited reflection and discussion regarding the risks of unfairness and specific measures to mitigate them during both the creation and utilisation of these tools. Laranjômetro, which scores the risk of straw candidacies, stands out as an exception, since its creators not only identified and tried to mitigate the issue of disproportionately targeting Black women from lower educational and social backgrounds as more likely to be straw candidates, but also openly documented it in a report.

The existing literature warns that, when approached without careful consideration, AI-based predictive law enforcement tools have the potential to reproduce existing patterns of discrimination and inherent biases from previous databases and decision makers (Barocas & Selbst, 2016; Richardson et al., 2019; Shapiro, 2017, 2019). These tools can also reflect the pervasive biases that exist in society at large (Edler Duarte, 2021). However, this study suggests that even when developers tend to be careful and try different sorts of techniques to compensate for eventual statistical issues, AI-ACTs can still exacerbate existing inequalities by suggesting that certain groups of people may be receiving less favourable treatment. This became clear in the case of Laranjômetro for Black women from lower social background, in the case of RCPC for recently created companies, and in the case of MARA for civil servants working in the units with a higher punishment record and affiliated to political parties. In the end, even if only used to guide human monitoring activities, these tools may contain implicit biases and reflect past injustices, as their algorithms are based on past anti-corruption procedures and practices that can be problematic. Findings also suggest that the AI-ACT developers of the tools analysed have not considered the potential risks of unfairness when creating their tools.

Although this study thoroughly explores the theme of (un)fairness in AI-based anti-corruption efforts in the Brazilian context, it also sheds light on the risks associated with the hype surrounding the use of technology to combat corruption everywhere, which often lacks in-depth debate. The limited number of studies assessing these types of tools suggest that we not only lack well-developed mechanisms for ensuring the integrity of existing predictive anti-corruption technologies but are also ill-prepared for the new challenges posed by generative systems. This is not to say that the overly tech-oriented anti-corruption policies cannot be implemented, but it seems urgent to have a better legal and methodological apparatus able to audit, identify, and correct unfairness issues.

If algorithmic biases and injustices are not properly addressed or mitigated, they can perpetuate unfairness and contribute to unequal treatment in both digital and

non-digital law enforcement, such as observed patterns in fields including policing violent crimes. In short, this study contributed to the existing literature by documenting the development of various AI-ACTs and highlighting their potential sources of unfairness. As a key takeaway, it underscores the need for more open discussions on AI fairness among academics and practitioners to identify potential areas of concern and the urgent need to assess existing AI-ACTs. The analytical framework for assessing potential sources of unfairness at the individual, institutional, or infrastructural levels can serve as a useful tool to pinpoint critical issues and provide clear guidelines for testing and mitigating unfairness, not only in ACTs. After all, it is not only the anti-corruption field that needs more debate on opening AI codes for inspection, promoting diversity among developers in terms of gender, race, and social background, and enhancing the quality of corruption data. By taking these steps, we can learn from past experiences without perpetuating historical unfairness, whether conscious or unconscious.

Acknowledgements The author acknowledges that the research for this article has been conducted as part of the BIT-ACT (Bottom-Up Initiatives and Anti-Corruption Technology) project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802362). Preliminary drafts of this monograph have been presented at the following international conferences: 2023 ECPR–European Consortium for Political Research General Conference (September 2023); 2024 IRSPM–International Research Society for Public Management (April 2024); 2024 LASA–Latin American Studies Association (June 2024), and the 3rd European Workshop on Algorithmic Fairness–EWAF'24 (July 2024). The author is grateful for the valuable feedback provided by those who read earlier versions of this article, including the anonymous reviewers. There are no conflicts of interest.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Aarvik, P. (2019). *Artificial Intelligence – a promising anti-corruption tool in development settings?* Report, U4, 1–38. Retrieved May 16, 2023, from https://www.u4.no/publications/artificial-intelligence-a-promising-anti-corruption-tool-in-development-settings

Adam, I., & Fazekas, M. (2021). Are emerging technologies helping win the fight against corruption? A review of the state of evidence. *Information Economics and Policy, 57*(100950), 1–14. https://doi.org/10.1016/j.infoecopol.2021.100950

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022). Contestable AI by design: Towards a framework. *Minds and Machines*. https://doi.org/10.1007/s11023-022-09611-z

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

16, 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aradau, C., & Blanke, T. (2017). Politics of prediction: Security and the time/space of governmentality in the age of big data. *European Journal of Social Theory, 20*(3), 373–391. https://doi.org/10.1177/1368431016667623

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(671), 671–732. https://doi.org/10.2139/ssrn.2477899

Beigang, F. (2022). On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. *Minds and Machines, 32*(4), 655–682. https://doi.org/10.1007/s11023-022-09615-9

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research, 81*, 1–11. https://doi.org/10.48550/arXiv.1712.03586

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*, pp. 405–415. Springer. https://doi.org/10.48550/arXiv.1707.01477

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. Sage Publications.

Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity, 56*, 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Carvalho, R.S. (2015a). *Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos Federais*. Universidade de Brasília. Retrieved May 25, 2023, from https://repositorio.unb.br/bitstream/10482/19361/1/2015_RicardoSilvaCarvalho.pdf

Carvalho, R.S. (2015b). *Filiação Partidária e Risco de Corrupção de Servidores Federais*. Retrieved May 25, 2023, from https://pt.slideshare.net/rommelnc/filiao-partidria-e-risco-de-corrupo-de-servidores-pblicos-federais

Carvalho, R.N. (2016). *Mapeamento de Risco de Corrupção na Administração Pública*. Retrieved May 25, 2023, from https://pt.slideshare.net/rommelnc/mapeamento-de-risco-de-corrupo-na-administrao-pblica-federal

Carvalho, R., Carvalho, R., Ladeira, M., Monteiro, F.M., & Mendes, O. (2014a). *Using political party affiliation data to measure civil servants' risk of corruption*. 2014 Brazilian Conference on Intelligent Systems, Sao Paulo, Brazil, 2014, pp. 166–171. https://doi.org/10.1109/ICPP.2014.39

Carvalho, R.N., Sales, L., Rocha, H.A., & Mendes, G.L. (2014b). *Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil*. Proceedings of the Eleventh UAI Bayesian Modeling Applications Workshop (BMAW 2014), Quebec, Canada, July 27, 2014. Retrieved May 25, 2023, from https://ceur-ws.org/Vol-1218/bmaw2014_paper_7.pdf

Center for Effective Global Action. (2018). *Machine Learning to Fight Corruption in Brazil -- Thiago Marzagão*. YouTube. Retrieved May 25, 2023, from https://www.youtube.com/watch?v=2prrNVaD-Nc

Ceva, E., & Jiménez, M. C. (2022). Automating anti-corruption? *Ethics and Information Technology, 24*, 48. https://doi.org/10.1007/s10676-022-09670-x

Cheibub, J. A., & Sin, G. (2020). Preference vote and intraparty competition in open list PR systems. *Journal of Theoretical Politics, 32*(1), 70–95. https://doi.org/10.1177/0951629819893024

Chen, S. (2019). *Is China's corruption-busting AI system 'Zero Trust' being turned off for being too efficient? South China Morning Post*. Retrieved September 29, 2022, from https://www.scmp.com/news/china/science/article/2184857/chinas-corruption-busting-ai-system-zero-trust-being-turned-being

Conaty, F. (2021). Abduction as a methodological approach to case study research in management accounting — An illustrative case. *Accounting, Finance & Governance Review*. https://doi.org/10.52399/001c.22171

Danks, D., & London, A.J. Algorithm bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence AI and autonomy track*, pp. 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Dincer, O., & Johnston, M. (2020). Legal corruption? *Public Choice, 184*, 219–233. https://doi.org/10.1007/s11127-020-00832-3

Domingos, S.L., Carvalho, R.N., Carvalho, R.S., & Ramos, G.N. (2016). *Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning*. Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 2016, pp. 722–727. https://ieeexplore.ieee.org/document/7838233

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*, eaao5580. https://doi.org/10.1126/sciadv.aao5580

Edler Duarte, D. (2021). The making of crime predictions: Sociotechnical assemblages and the controversies of governing future crime. *Surveillance & Society, 19*(2), 199–215. https://doi.org/10.24908/ss.v19i2.14261

Eidelson, B. (2015). *Discrimination and disrespect*. Oxford University Press.

Compartilhado, G. (2020). *Laranjômetro-Previsão de candidaturas laranjas em 2020* - Nota Técnica nº 003/2020. Retrieved May 16, 2023, from https://henriquexavier.net/laranjometro/nota_tecnica_laranjometro_2020.pdf

Grunewald, C. & Cosac, D. (2016). *Risk of public contracts: machine learning + multi-criteria decision analysis*. Retrieved May 28, 2023, from https://pt.slideshare.net/rommelnc/proposta-de-modelo-de-classificao-de-riscos-de-contratos-pblicos

Halpern, C., & Le Galès, P. (2011). No autonomous public policy without ad hoc instruments: A comparative and longitudinal analysis of the European Union's environmental and urban policies. *Revue Française De Science Politique, 61*, 51–78. https://doi.org/10.3917/rfsp.611.0051

Jambreiro Filho, J. (2019). *Artificial Intelligence Initiatives in the Special Secretariat of Federal Revenue of Brazil*. Retrieved September 29, 2022, from www.jambeiro.com.br/jorgefilho/AI_Brazil_Federal%20Revenue%20_2019.pdf.

Jefferson, B. J. (2018). Predictable policing: Predictive crime mapping and geographies of policing and race. *Annals of the American Association of Geographers, 108*(1), 1–16. https://doi.org/10.1080/24694452.2017.1293500

Johnston, M. (2005). *Syndromes of corruption: Wealth, power, and democracy*. Cambridge University Press.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2012). Techniques for discrimination-free predictive models. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society* (pp. 223–240). Springer.

Kahneman, D., Sibony, O., & Sustein, C. R. (2021). *Noise – A flaw in human judgement*. Little, Brown Spark.

Köbis, N., Starke, C., & Rahwan, I. (2022). The promise and perils of using artificial intelligence to fight corruption. *Nature Machine Intelligence, 4*, 418–424. https://doi.org/10.1038/s42256-022-00489-1

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*. Retrieved May 28, 2023, from https://doi.org/10.48550/arXiv.2102.11567

Kossow, N., Windwehr, S., & Jenkins, M. (2021). *Algorithmic transparency and accountability. Transparency International Anti-Corruption Helpdesk Answer*. Retrieved May 28, 2023, from https://knowledgehub.transparency.org/assets/uploads/kproducts/Algorithmic-Transparency_2021.pdf

Lagunes, P., Michener, M., Odilla, F., & Pires, B. (2021). President bolsonaro's promises and actions on corruption control. *Revista Direito GV*. https://doi.org/10.1590/2317-6172202121

Lascoumes, P., & Le Galès, P. (2007). Introduction: understanding public policy through its instruments - from the nature of instruments to the sociology of public policy instrumentation. *Governance, 20*(1), 1–21. https://doi.org/10.1111/j.1468-0491.2007.00342.x

Lessig, L. (2013). "Institutional corruption" defined. *Journal of Law, Medicine & Ethics, 41*(3), 553–555. https://doi.org/10.1111/jlme.12063

Lima, O.D.W., & Andrade, N. (2019). *Fairness in Risk Estimation of Brazilian Public Contracts.* Symposium on Knowledge Discovery, Mining and Learning, KDMILE 2019 - Applications Track. Retrieved May 28, 2023, from https://sol.sbc.org.br/index.php/kdmile/article/view/8789/8690. Retrieved 28 May 2023.

Lima, O.D.W. (2020). *Justiça em aprendizagem de máquina na estimativa de risco de contratos públicos*. Universidade Federal de Campina Grande. Retrieved May 28, 2023, from http://dspace.sti.ufcg.edu.br:8080/jspui/bitstream/riufcg/15996/1/%C3%93RION%20DARSHAN%20WINTER%20DE%20LIMA%20%E2%80%93%20DISSERTA%C3%87%C3%83O%20%28PPGCC%29%202020.pdf

Marzagão, T. (2017). Using AI to fight corruption in the Brazilian government. Retrieved May 28, 2023, from https://speakerdeck.com/thiagomarzagao/using-ai-to-fight-corruption-in-the-brazilian-government

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 2021*(8), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Moreau, S. (2010). What is discrimination? *Philosophy & Public Affairs, 38*, 143–179. https://doi.org/10.1111/j.1088-4963.2010.01181.x

Odilla, F. (2020). Oversee and Punish: Understanding the Fight Against Corruption Involving Government Workers in Brazil. *Politics and Governance, 8*, 2. https://doi.org/10.17645/pag.v8i2.2716

Odilla, F. (2022). *Avoiding minority reports: using AI responsibly in anti-corruption*. Corruption in Fragile States Blog. Retrieved May 18, 2023, from https://www.corruptionjusticeandlegitimacy.org/post/avoiding-minority-reports-using-ai-responsibly-in-anti-corruption

Odilla, F. (2023). Bots against corruption: Exploring the benefits and limitations of AI-based anti-corruption technology. *Crime Law and Social Change*. https://doi.org/10.1007/s10611-023-10091-0

Odilla, F. (2024). *The digitalisation of anti-corruption in Brazil: Scandals, reforms, and innovation*. Routledge.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., & Oliveira, E. L. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing, 2023*(7), 15. https://doi.org/10.3390/bdcc7010015

Poltoratskaia, V., & Fazekas, M. (2023). *Corruption risk assessments: country case studies highlight advantages and challenges of diverse approaches*. U4 Anti-Corruption Resource Centre, Chr. Michelsen Institute (U4 Issue 2023:2). Retrieved June 8, 2023, from https://www.u4.no/publications/corruption-risk-assessments-country-case-studies-highlight-advantages-and-challenges-of-diverse-approaches

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 94, pp. 15–55. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

Silva, S., & Kenny, M. (2018). Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. *Phylon: The Clark Atlanta University Review of Race and Culture*, (Summer/Winter 2018) Vol. 55, No. 1 & 2: pp. 9–37. https://www.jstor.org/stable/26545017

Saleiro, P., Benedict Kuester, A. S., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). *Aequitas: A bias and fairness audit toolkit*. https://doi.org/10.48550/arXiv.1811.05577

Sales, L.J. (2016). *Proposta de modelo de classificação do risco de contratos públicos.* Universidade de Brasília. Retrieved May 28, 2023, from http://mesp.unb.br/images/dissertacaoes/2016/Dissertacao-de-mestrado---Leonardo-Sales.pdf.

Santiso, C. (2019, February 28). *Here's how technology is changing the corruption game.* World Economic Forum. Retrieved May 23, 2023, from https://www.weforum.org/agenda/2019/02/here-s-how-technology-is-changing-the-corruption-game/

Shapiro, A. (2017). The medium is the mob. *Media, Culture & Society, 39*(6), 930–941. https://doi.org/10.1177/0163443717692740

Shapiro, A. (2019). Predictive policing for reform? indeterminacy and intervention in big data policing. *Surveillance & Society, 17*(3/4), 456–472. https://doi.org/10.24908/ss.v17i3/4.10410

Sharma, V. (2018, November 15). *Can artificial intelligence stop corruption in its tracks?* World Bank Blogs. Retrieved May 27, 2023, from https://blogs.worldbank.org/governance/can-artificial-intelligence-stop-corruption-its-tracks

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Siegel, E. (2018). *How to fight bias with predictive policing*. Scientific American Blog. Retrieved May 23, 2023, from https://blogs.scientificamerican.com/voices/how-to-fight-bias-with-predictive-policing/.

Singh, J. P., Desmarais, S., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law, 31*(1), 55–73. https://doi.org/10.1002/bsl.2053

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* July 2019, pp. 2459–2468. https://doi.org/10.48550/arXiv.1902.04783

Starke, C., Kieslich, K., Reichert, M., & Köbis, N. (2023). *Algorithms against Corruption: A conjoint study on designing automated twitter posts to encourage collective action*. https://doi.org/10.31235/osf.io/wf45t

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*. https://doi.org/10.1177/20539517221115189

Sun, T., & Sales, L. J. (2018). Predicting public procurement irregularity: An application of neural networks. *Journal of Emerging Technologies in Accounting, 15*(1), 141–154. https://doi.org/10.2308/jeta-52086

TCU. (2022). *Acórdão 1139/2022. Pesquisa Integrada do TCU*. Retrieved May 30, 2023, from https://pesquisa.apps.tcu.gov.br/#/documento/acordao-completo/*/NUMACORDAO%253A1139%2520ANOACORDAO%253A2022/DTRELEVANCIA%2520desc%252C%2520NUMACORDAOINT%2520desc/0/%2520

Turtelli, C. (2020). *Estudo indica ao menos 5 mil candidatas laranjas nas eleições 2020*. Estado de S.Paulo. Retrieved May 30, 2023, from https://www.estadao.com.br/politica/eleicoes/estudo-indica-ao-menos-5-mil-candidatas-laranjas-nas-eleicoes-2020/

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security, 23*(4), 323–340. https://doi.org/10.1177/13882627211031257

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*. https://doi.org/10.1177/2053951717743530

Xavier, H. (n.d.). *Laranjometro*. Retrieved May 16, 2023, from https://henriquexavier.net/laranjometro.html

Zehlike, M., et al. (2017). FA*IR: A fair top-k ranking algorithm. *CIKM'17,* November 6–10, 2017, Singapore. https://doi.org/10.48550/arXiv.1706.06368