The Promise and Perils of using Artificial Intelligence to fight Corruption

Nils Köbis[1], Christopher Starke[2] & Iyad Rahwan[1]

[1] Center for Humans and Machines, Max-Planck-Institute for Human Development

[2] Human(e) AI, Amsterdam School of Communication Research, University of Amsterdam

Correspondence to: Nils Köbis, koebis@mpib-berlin.mpg.de, Max-Planck-Institute for Human Development, Center for Humans and Machines, Lentzallee 94, 14195 Berlin, Tel.: +49 30 82406-751/752

Abstract

Corruption presents one of the biggest challenges of our time, and much hope is placed in Artificial Intelligence (AI) to combat it. While the growing number of AI-based anti-corruption tools (AI-ACT) have been summarised, a critical examination of their promises and perils is lacking. Here, we argue that the success of AI-ACT strongly depends on whether they are implemented top-down (by governments) or bottom-up (by citizens, NGOs, or journalists). Top-down use of AI-ACT can consolidate power structures and thereby pose new corruption risks. Bottom-up use of AI-ACT has the potential to provide unprecedented means for the citizenry to keep their government and bureaucratic officials in check. We outline the societal and technical challenges that need to be overcome to harness the potential for AI to fight corruption.

Keywords: Anti-corruption; Digital Technologies; Open Government Data; Artificial Intelligence; Power

Corruption – commonly defined as the abuse of entrusted power for private gains[1,2] – presents one of the biggest societal and political challenges[3–5]. While vast (financial) efforts have been invested in the fight against corruption, they have shown little signs of success[6,7]. Advancements in the field of Artificial Intelligence (AI), defined here as "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals"[8] provide a new beacon of hope. Using AI technologies in the fight against corruption can bring a long-awaited transformative change, so the aspirations read. Indeed, already praised as "the next frontier in anti-corruption"[9], governments, donor organisations, and non-governmental organisations (NGOs) have begun implementing AI technologies in anti-corruption efforts[10,11].

In this Perspective, we (a) summarize the main reasons for hope of AI technologies positively transforming anti-corruption efforts, (b) highlight challenges to be met and (c) draw attention to the diverging effects of AI technologies on existing power structures when used top-down (e.g. by governments) versus bottom-up (e.g. by civil society organizations).

AI-based anti-corruption tools, labelled AI-ACT henceforth, are appealing for fighting public corruption (see Box 1 for an explanation of different corruption types). Namely, AI has three key advantages over traditional anti-corruption efforts. First, AI systems, such as machine-learning, can be imbued with **autonomous learning abilities.** Hence, unlike static information communication technology (ICT), AI can independently execute various tasks previously reserved to human actors[12]. Many researchers and policymakers expect these autonomous learning abilities to detect and even predict corruption (risks)[13,14]. Indeed, innovative projects showcase that AI-ACT can learn and automatically detect risk zones for corruption[15] and already use a large corpus of news media reports[16], police archives[17], and data from financial authorities[15] to predict embezzlement or bribery.

Box 1 - Distinction between Public and Private Corruption Types

Corruption is an umbrella term that encompasses many different behaviors[18,19]. Successful anti-corruption requires specifying the respective type of corruption at hand[20]. One main distinction pertains to whether corruption occurs in the private or the public sector.

Private corruption refers to abuses of power not entrusted within the public sector, such as embezzlement by managers or bribe payments in business-to-business transactions.

Public corruption refers to abuses of entrusted power for private gains in the public sector. For example, public corruption ranges from heads of states embezzling public funds to lower-ranking public officials like traffic police officers requesting bribes. This *Perspective* focuses

on public corruption as most tools have been developed to tackle it, and it arguably presents the more harmful type of corruption for society.

Second, thanks to growing **computing power**, AI can analyse data sets of unprecedented size. This computational ability plays a crucial role in keeping track of newly emerging complex corruption schemes, such as kleptocrats using intricate webs of shell companies to hide their ill-gotten gains[21]. In an increasingly digitised world, fighting crime like corruption turns into an arms race of technology[22,23] and thus, the demand for AI tools to fight corruption grows[14,24]. Fuelled on a growing body of available data[14,24], AI systems can help classify and detect corrupt activities. For instance, Microsoft recently announced its Artificial Intelligence Technology Solutions project offering its anti-corruption products to governments[25]. AI tools can also sift through large data leaks such as the Pandora Papers to unveil corrupt patterns[25] – a task infeasible for humans alone[26].

Third, AI is, in principle, **impartial**. Human decision-makers, in particular, those holding public offices, often face conflicts of interests. Research in behavioural science has provided ample evidence that humans tend to bend moral rules for their benefit, especially when (financial) temptations exist[meta-analysis: 27] or under time pressure[meta-analysis: 28]. Algorithms, however, pursue no self-serving interests and process information in a "disinterested" way[29]. They furthermore make decisions consistently, unaffected by time pressure or fatigue[30]. Therefore, it is appealing to replace human decision-makers with AI, especially in contexts where corruption is widespread. In such contexts, those actors tasked with fighting corruption often fall prey to high levels of corruption themselves: a so-called corruption trap emerges[3,31]. It ranges from police officers halting investigations for bribes to prosecutor generals selectively pressing charges based on political agendas. AI has unprecedented potential to help escape this corruption trap. Namely, with no human-in-the-loop[32], the process of AI-ACT cannot be tampered with by human decision-makers.

A burgeoning hype kill?

Yet, signs of a potential hype kill are already in sight. Interest in AI has experienced several "AI winters" when technological advancements could not live up to the high expectations[33]. And such could become the case for AI in the fight against corruption. Besides this general risk of reality and expectations diverging, three unique challenges exist for AI-ACT.

The first one is a **data challenge**. Obtaining valid and reliable data to establish a ground truth presents a particularly thorny challenge for corruption. Extensive research has debated what indicators serve as valid proxies for corruption and how to measure a phenomenon that is, per definition, hidden from plain sight[34]. An illustrative example stems from Brazil, where the project MARA uses machine learning to calculate an individual-level corruption score based on previous conviction data[35]. Training a machine-learning algorithm on past convictions pays dividends when authorities impartially sanction corrupt practices. However, such data sources typically suffer from biases, for instance, when political agendas drive investigations, prosecutions, and convictions of corruption[35]. Similarly, media reports about corruption often reflect the journalistic quality or media freedom rather than an accurate representation of corruption occurrences in the respective country[36]. Documenting and eventually reducing such biases requires an observable ground truth, which is difficult to establish for corruption.

The second is an **algorithm challenge**. Choices made by intelligent algorithms often have far-reaching, value-laden consequences[36,37], especially when it comes to corruption[36,37]. As with any classification algorithm, AI systems seeking to categorise cases into "true" versus "false" face a trade-off between false-positive and false-negative errors[37,38]. For AI-ACT, false-positive errors, the wrong classification of innocent individuals as "corrupt", come with a particularly strong stigma[39]. Once publicised, those accused of corruption tend to be prematurely prosecuted in the court of public opinion. They, thus, suffer irreversible reputation losses. Wrongful accusations might also reduce citizens' trust in such AI-ACT. Conversely, false-negative errors come with the cost of leaving actual corrupt cases undetected, which means non-negligible costs for public institutions or society[39,40]. Just consider if an algorithm employed by the government fails to detect blatant corruption cases. The public might quickly turn against it, suspecting that those in power tinker with the AI system. Consequently, avoiding backlash against AI-ACT requires a balancing act of minimising false-positive versus false-negative error rates.

Third, a **human challenge** exists. Implementing AI-ACT is not a trivial task because algorithms never operate in a vacuum but are embedded in socio-institutional contexts[39,40]. Harnessing the potential of AI-ACT requires setting a suitable degree of autonomy yielded to the algorithms: A tension exists between the ethical principle of keeping humans in control of decisions[40,41] versus letting AI decide autonomously to escape the aforementioned corruption trap. In all public domains where AI makes societally relevant decisions, such transfer of

decision authority from humans to algorithms requires legitimacy. Wielding autonomy to AI removes established checks and balances, which increases the risk of harmful outcomes such as false corruption accusations. Consequently, people often distrust AI to make final decisions in the public domain[42]. Also, for those humans working in hybrid teams with AI systems, relying on AI can lead to "algorithmic dumbfounding"[43] as humans might blindly follow algorithmic recommendations.

Unique about AI-ACT is that delegating responsibility to the AI system might lead to human-decision makers losing their ability to make decisions adequately and lacking the holistic aspects of the task[44]. For example, when corruption detection becomes increasingly AI-based, humans might lose the skill of sniffing out a bribe offer based on indirect speech acts[45]. Another human risk of AI-ACT is that public officials might feel less responsible for engaging in anti-corruption activities like whistleblowing themselves.

AI for top-down versus bottom-up and its different effects on power structures

The promises and perils outlined so far apply to both government-led top-down and citizen-led bottom-up anti-corruption approaches (see for more details on both types of approaches Fig. 1). However, a significant difference emerges when analysing the use of AI systems in the fight against corruption through a social science lens that puts power centre stage[46]. AI-ACT used in top-down efforts risks reinforcing existing power structures, whereas AI-ACT used in bottom-up efforts might help shift them.

When it comes to the excitement around top-down use of AI-ACT, a widely unacknowledged concern deserves attention: using AI-ACT top-down can lead to a consolidation of power, which introduces new (corruption) risks. Be it governments or companies, power rests with those who have data and code[47–49]. While AI-ACT are set up to mitigate the harm of corruption, they are run by powerful institutions that pursue their own agendas[47], such as governments seeking to remain in office.

A characteristic of top-down AI-ACT is the technical infrastructure of access and aggregation of (sensitive) data combined with powerful algorithms enabling unprecedented surveillance and control[48,49]. Commercial providers explicitly promote their top-down AI-ACT services to governments as tools for "more effective controls"[50]. The "Zero Trust" project introduced by the Chinese government to stamp out corruption among its workforce of over 60 million public officials allows a glimpse into such tools in action[11,51]. It employed AI tools that

cross-referenced 150 protected databases, featuring public officials' bank statements, property transfers, and private purchases, to calculate probabilities of corrupt activities. This example illustrates that access and aggregation of (private) databases facilitate the pursuit of corruption – which is true for most top-down AI-ACT.

Therefore, the technical infrastructure of AI-ACT provides new means of consolidating power, which poses new risks of abuse. Extensive empirical evidence across disciplines illustrates that concentration of power tends to breed power abuses[52,53, review: 54] – encapsulated in the famous adage that "power corrupts; absolute power corrupts absolutely"[55]. So-called "Big Brother Effects" can emerge too[44], where governments use AI tools to monitor and weave out political opposition[44]. Indeed, econometric evidence supports the corruption risks of such digital surveillance infrastructure[56]. Hence, under the guise of fighting corruption, governments (and companies) might use AI to consolidate power and undermine instead of strengthening democratic institutions.

Such risks are particularly pressing in socio-economic contexts characterised by a weak rule of law[56]. Here, AI-based anti-corruption campaigns "are primarily used by incumbents to target political opponents – algorithms may simply help governments to crack down on critics more efficiently"[57]. One of the biggest challenges for anti-corruption efforts in corrupt contexts is that those in power abusing it for private gains have little incentive to change the power structure and reduce corruption[31]. Especially when the rule of law is weak, AI systems tend to reinforce societal power structures[47], new risks for power abuses emerge, and gaps in inequalities widen.

Are AI-ACT in bottom-up efforts flipping the script?

Although several reasons for scepticism exist when governments employ AI-ACT top-down, more optimism is warranted about how AI-ACT can support bottom-up initiatives[58]. Bottom-up approaches seek to reduce corrupt practices by analysing the given socio-cultural context and support existing efforts by civil society organisations, NGOs, and investigative journalists[36,59,60]. Enabling protest and other forms of collective action are crucial for democratic regimes to emerge[61] and corruption to diminish[62].

In contrast to top-down AI-ACT, where governments scrutinise public servants and citizens[63–65], bottom-up AI-ACT can flip the script. They allow citizens to organise better and scrutinise their government (officials). Here lies a unique potential of AI-ACT. Instead of the

government taking the role of a Big Brother, AI-ACT used in bottom-up efforts can allow the public to turn into watchdogs, keeping the government in check.

Promising cases of AI systems assisting bottom-up approaches to fight corruption are starting to appear[64]. In Ukraine, the portal Dozorro draws on AI to flag public procurement tenders with high corruption risks and communicate them to the public[66]. In Brazil, the Tweetbot "Rosie da Serenata" automatically analyses publicly available government data on reimbursement claims of government officials and autonomously detects suspicious cases[67]. It tweets such cases out and encourages its followers to investigate them further[35,65]. In Nigeria, the DataCrowd project applies AI technology like computer vision and, in the future, Natural Language Processing (NLP) to enable citizens to monitor public projects and reduce corruption[24,36].

How to increase the success of bottom-up AI-ACT?

Achieving the potential of bottom-up AI-ACT to shift power structures and reduce corruption requires **data to fuel the algorithms**. Public administration around the world is becoming increasingly digital[68]. E-government initiatives, open data programs, and citizen-driven crowdsourcing efforts render more data publicly available[68,69]. While this is a laudable trend, the vast majority of data remains undisclosed and in the hands of governments or companies. This lack of available data hinders bottom-up AI-ACT from unleashing their potential.

Data sources, currently not used to fight corruption, could be employed for bottom-up AI-ACT in the future. Consider so-called data traces[70]. Digital technologies like smartphone apps or sensors embedded in people's daily lives collect and store traces of human behaviour. Such digital data traces feature social media communication, geospatial data, browser history, and contextual data about when, where, and how behaviour occurs[71].

First legal frameworks (e.g., the European Union's General Data Protection Regulation) mandate that digital platforms provide users with a copy of their data and allow them to access it via "data download packages" (DDPs). Encouraging people to "donate" such DDPs is a growing trend in social science research[72] and could also enable AI-ACT initiatives. For example, data traces that reveal people's geospatial movement patterns could help identify flaws in public infrastructure that can hint at corrupt public construction processes[similar project: ,73].

Future apps could also automatically log interactions between citizens and public officials to document cases or risks of corruption. Such efforts could scaffold on existing apps like "Siri, I'm being pulled over," which automatically records citizens' encounters with police officers[74]. Soon, the scope of such apps could be extended to other public domains.

Such efforts could provide valuable data, particularly for bottom-up efforts, to train algorithms about the predictors of corruption, including specific offices, regions, and sectors. Using data traces should be accompanied by a broad public discussion about which data sources should be used and which ones should be off-limits. Research assessing people's views about the emerging moral trade-offs between fighting corruption and infringing on people's privacy can aid such efforts.

A second requirement is to **foster and sustain collective action**. Successful collective action, in general, requires the mobilisation and sustained engagement of citizens[75]. Fighting corruption further involves the promotion of not only transparency but also accountability. It has long been assumed that the growing availability of information will enable citizens to educate and coordinate themselves in the fight against corruption[76]. Yet, transparency alone does not suffice to curb corruption[77,78]. Whether prosecutors, journalists, or civil society actors, someone needs to draw inferences from data to render it actionable for policy efforts[79]. Transparency on paper needs to be turned into action to advance accountability[78], as transparency without accountability is like the "sound of one hand clapping"[80].

AI-ACT can help to foster both transparency and accountability. Traditional – non-AI-based – collective action efforts like crowdsourcing platforms already hint at the immense potential to promote transparency with the help of technology. Digital crowdsourcing tools have enabled citizens to report many corruption cases[81]. For example, via the Trade Route Incident Mapping System (TRIMS) in Nigeria[82,] truckers and small traders could use their phones to report when and where they were extorted to pay bribes in traffic checkpoints.

AI tools could help such transparency efforts by facilitating the reporting of corruption cases. Extending previous tools that used written corruption reports, AI-based efforts could draw on chatbots or voice-bots that ask about the crucial aspects of the case. This AI integration, in turn, could lower the initial threshold to report and render the collected data more useful. Other multimedia inputs are possible too. In Mexico, a smartphone app enabled citizens to document shortcomings in public infrastructure by taking short videos and geo-tag them[83].

NLP-technology and classification algorithms could automatically extract opinions, sentiments, or geospatial patterns from such reporting and extract relevant information for policy efforts. Such functionality has been piloted in a project on AI-assisted citizen engagement in Nigeria[73]. Besides motivating people to report corruption, AI-ACT can autonomously do the reporting itself, as is already the case for the Tweetbot "Rosie da Serenata" that automatically tweets about suspicious expense claims by Brazilian parliamentarians.

Such collective action efforts need to be sustained to have a lasting impact. Here, the non-AI-assisted reporting efforts provide a warning sign. The engagement often fizzles and, in many cases, dies out altogether. A common difficulty for such collective action efforts is to go beyond the initial mobilisation phase and keep people engaged. As a case in point, the website of the TRIMS project is no longer accessible.

AI systems could elevate such past efforts and turn the initial enthusiasm of crowdsourced reporting projects into more sustained efforts. AI agents have already become part of online (political) communities[84]. For collective anti-corruption efforts, AI agents could particularly take the role of a "dedicated motivator" who keeps others engaged – a crucial part of any social movement[75]. For instance, it could send personalised messages tailored to each citizen based on previous activities to motivate re-engagement. Such personalised messaging has already shown first success in other crowd-civic efforts[85].

Finally, beyond mobilising citizens to report corruption and fostering transparency, AI-ACT can also spur accountability. Pilot projects on Twitter provide a blueprint for such efforts seeking to animate people to participate in activism against corruption. AI text-classifiers can detect posts about corruption on Twitter, which helps to identify those with interest in (anti-)corruption[86]. One step further, the Botivist project programmed a Tweetbot that contacted people tweeting about corruption and impunity[87]. The conversational agent then encouraged them to engage in collective action against corruption, such as signing petitions. It also invited Twitter users to brainstorm solutions for corruption and suggested collaborations, thereby facilitating social activism. Overall, the bot achieved a 45% response rate hinting at the potential of using AI to mobilise citizens for collective action against corruption.

Akin to top-down implementation, the success of bottom-up AI-ACT depends on the socio-economic context. Digital collective action typically requires a smartphone, internet access, and technical skills – prerequisites unevenly distributed within and across societies

around the world[36]. Furthermore, such technology-based anti-corruption efforts flourish in countries with high media freedom[87] and freedom of expression.

Future Scenario: Coupling AI-ACT with other technologies

A big upside exists in coupling AI-ACT with other digital technologies, especially those that can offset AI's limitations in transparency and privacy. As outlined, AI-ACT requires access to or publication of (private) data. However, such data transparency can pose a risk to individual privacy. For instance, unmasked data leaks, such as some data exposed on Wikileaks, neglected privacy concerns[88]. Hence, for AI-ACT, a trade-off between transparency and privacy emerges[89]. Distributed ledger technologies (DLT) such as blockchain can alleviate this tension.

DLT are data storage systems that use peer-to-peer networks of independent nodes[90]. Every network node stores an identical copy of the database. Since all entities such as individual people, companies, or institutions like NGOs can operate a node, they secure the network[91]. Nodes independently validate transactions on the network through an algorithmic con-sensus mechanism. For example, "Proof of Work" proves that nodes expended computational energy to validate transactions. These transactions are then stored in a timestamped chain of blocks: a blockchain. Through cryptography, the records are immutable and cannot be tampered with by malicious actors[90,91]. By pseudonymising transactions and publicising records, DLT can contribute to privacy[92] and transparency[93].

In *permissioned* DLT, nodes are operated by centralised entities. Such technologies have successfully been implemented top-down in public administration[94,95, review: 96], inter alia to reduce corruption[97]. For example, the government of Georgia has transferred its land registry into a permissioned DLT[98]. This move brings with it benefits for AI-ACT. It provides a secure database with timestamped transfers of records that enable immutable proof of ownership. Such high-quality data, transparently available and privacy-preserving, is the perfect fuel for AI-ACT to detect patterns of corruption. However, such implementation of DLT still hinges on the quality of the institutions that implement it. Namely, permissioned DLT are still prone to corruption risks based on transaction forgery, data manipulation, and censorship stemming from the centralised institution's opportunistic behavior[99].

Let us sketch a future scenario in which *permissionless* DLT and AI join forces to unleash the full potential of confronting corruption with digital technologies. For starters, DLT could help achieve an age-old goal of anti-corruption efforts: "follow the money". Consider

that public procurement accounts for 30 to 50 per cent of public spending globally[100], amounting to approximately $11 trillion awarded in government contracts annually[101]. Corrupt activities divert a sizable portion of that public money into private pockets. Imagine a government introducing a blockchain-based public expenditure tracking system. Like already existing DLT projects that track supply chain management in the private sector[97], future DLT could store all public procurement expenditures and related sub-contracting transactions. The government pays a small fee to incentivise node operators to store transaction data by governments and companies in a transparent, tamper-proof database[102]. Average citizens can become network nodes and are empowered to secure the database. AI tools, in turn, can autonomously audit the downstream transaction flows for corruption. One possibility for new forms of AI-based accountability is that after completing an audit of public procurement contracts, AI-ACT rewards the involved public officials with "integrity tokens" for delivering public services as promised.

Current-day technologies hint at the potential to create such possible futures. It is largely a policy decision whether they will come to fruition. Here, (funding for) AI research plays a crucial role towards such seemingly utopian versions of the future[44]. Whether AI systems can deliver on the hope of reducing corruption depends on developing such technologies. From an economic perspective, an essential question is whether economic demand for AI systems to monitor citizens supersedes the incentives for tools to empower citizens and facilitate collective action. Typically, such market demands are transmitted to AI researchers who thus either work on AI systems to consolidate existing power structures or help shift them towards the citizenry[32]. Funding for research projects on AI tools for bottom-up efforts to fight corruption thus marks an essential proximate step.

Conclusion

Already lauded as the next frontier in anti-corruption, using AI technologies to curb corruption is still in its infancy. Therefore, nascent decisions about using AI-ACT will shape how it affects (future) societies. We argue that top-down implementation of AI-ACT tends to consolidate existing power structures and, in turn, creates new risks of power abuses. This risk is especially prevalent in contexts where corruption is the rule rather than the exception. Here, top-down AI-based anti-corruption tools can be misappropriated by governments to enhance digital surveillance, suppress opposition, and undermine democratic liberties. In highlighting these perils, we argue that top-down AI-ACT must be introduced with extra caution. A concrete

recommendation for responsible top-down implementation of AI-ACT consists of ensuring active involvement of all relevant societal stakeholders, hence, having "society in the loop"[103]. Such a citizen-centred approach reduces the risks of abuse and heightens the legitimacy of such tools.

Bottom-up efforts take citizens' interests as their starting point. This starting point does not guarantee that these citizen-driven efforts succeed or are necessarily legitimate. Just consider that the advent of social media was similarly met with excitement for its promised democratising role[103]. Along these lines, bottom-up efforts using AI-ACT might go awry, such as programs to report corruption could lead to denunciation campaigns. Yet, when implemented responsibly, AI-ACT has the potential to become a driving force to mobilise previously apathetic citizens into new efforts to keep power holders accountable. Even in contexts where corruption has become endemic, citizens, NGOs, and journalists courageously engage in collective action against corruption. Supporting such efforts with AI-ACT could boost their chances of success. We hope that more AI developers join the anti-corruption community to explore the potential of enriching existing bottom-up initiatives with the power of AI technology.

Author contributions:

NCK: Conceptualisation, Writing original draft, Visualisation, Writing- Reviewing and Editing

CS: Conceptualisation, Writing original draft, Writing- Reviewing and Editing

IR: Writing- Reviewing and Editing, Supervision, Visualisation

Competing Interests Statement:

The authors declare no conflict of interests.

References

1. Rothstein, B. & Varraich, A. *Making Sense of Corruption*. (Cambridge University Press, 2017).

2. Köbis, N.C., van Prooijen, J.-W., Righetti, F. & Van Lange, P.A.M. The Road to Bribery and Corruption: Slippery Slope or Steep Cliff? *Psychol. Sci.* **28**, 297–306 (2017).

3. Fisman, R. & Golden, M.A. *Corruption: What Everyone Needs to Know*. (Oxford University Press, 2017).

4. Rothstein, B. *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective*. (University of Chicago Press, 2011).

5. Mungiu-Pippidi, A. & Heywood, P. *A Research Agenda for Studies of Corruption*. (Edward Elgar, 2020).

6. Mungiu-Pippidi, A. The time has come for evidence-based anticorruption. *Nature Human Behaviour* **1**, 0011 (2017).

7. Fisman, R. & Golden, M. How to fight corruption. *Science* **356**, 803–804 (2017).

8. High-Level Expert Group on Artificial Intelligence. *A Definition of AI: Main Capabilities and Disciplines*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341 (2019).

9. Petheram, A. From open data to artificial intelligence: the next frontier in anti-corruption. https://www.oxfordinsights.com/insights/aiforanticorruption (2018).

10. Aarvik, P. *Artificial Intelligence a promising anticorruption tool in development settings*. (2019).

11. World Bank. *Artificial intelligence in the public sector: Maximizing opportunities, managing risks*. (World Bank, 2020).

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

13. Mattoni, A. The grounded theory method to study data-enabled activism against corruption:

Between global communicative infrastructures and local activists' experiences of big data. *Eur. J. Disord. Commun.* **35**, 265–277 (2020).

14. Adam, I. & Fazekas, M. *Are emerging technologies helping win the fight against corruption in developing countries?* (2018).

15. Lavigne, S., Clifton, B. & Tseng, F. Predicting Financial Crime: Augmenting the Predictive Policing Arsenal. *Preprint at http://arXiv.org/abs/1704.07826* (2017).

16. López-Iturriaga, F.J. & Sanz, I.P. Predicting public corruption with neural networks: An analysis of spanish provinces. *Soc. Indic. Res.* **140**, 975–998 (2018).

17. De Blasio, G. *et al.* Predicting corruption crimes with machine learning. A study for the Italian municipalities. *Preprint at http://www.diss.uniroma1.it/sites/default/files/allegati/DiSSE_deBlasioetal_wp16_2020.pdf* (2020).

18. Lange, D.A. Multidimensional Conceptualization of Organizational Corruption Control. *AMRO* **33**, 710–729 (2008).

19. Heidenheimer, A.J. & Johnston, M. *Political Corruption: Concepts and Contexts*. (Transaction Publishers, 2011).

20. Heywood, P. Rethinking Corruption: Hocus-Pocus, Locus and Focus. *Slav. East Eur. Rev.* **95**, 21–48 (2017).

21. Obermaier, F. & Obermayer, B. *The Panama Papers: Breaking the Story of How the Rich and Powerful Hide Their Money*. (Simon and Schuster, 2017).

22. King, T.C., Aggarwal, N., Taddeo, M. & Floridi, L. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Sci. Eng. Ethics* **26**, 89–120 (2020).

23. Transparency International. Facing future corruption challenges — trends of the next decade. *Transparency.org* https://www.transparency.org/en/blog/facing-future-corruption-challenges-trends-of-the-next-decade (2019).

24. Kossow, N. Digital anti-corruption: hopes and challenges. in *A Research Agenda for Studies of Corruption* (eds. Mungiu-Pippidi, A. & Heywood, P.) 146–157 (Edward Elgar Publishing, 2020).

25. Anti-corruption technology solutions - about. https://www.microsoft.com/en-us/microsoftacts/about.

26. How can we use artificial intelligence to help us fight corruption in the mining sector? *Global Witness* https://www.globalwitness.org/en/blog/how-can-we-use-artificial-intelligence-help-us-fight-corruption-mining-sector/.

27. Leib, M., Köbis, N.C., Soraperra, I., Weisel, O. & Shalvi, S. Collaborative Dishonesty: A Meta-Study. *Psychol. Bull.* (2022).

28. Köbis, N.C., Verschuere, B., Bereby-Meyer, Y., Rand, D. & Shalvi, S. Intuitive Honesty Versus Dishonesty: Meta-Analytic Evidence. *Perspect. Psychol. Sci.* **14**, 778–796 (2019).

29. Giubilini, A. & Savulescu, J. The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philos. Technol.* **31**, 169–188 (2018).

30. Kahneman, D., Sibony, O. & Sunstein, C.R. *Noise: A Flaw in Human Judgment*. (Hachette UK, 2021).

31. Stephenson, M. Corruption as a Self-Reinforcing Trap: Implications for Reform Strategy. *World Bank Res. Obs.* **35**, 192–226 (2020).

32. Rahwan, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf. Technol.* **20**, 5–14 (2018).

33. Kerr, A., Barry, M. & Kelleher, J.D. Expectations of artificial intelligence and the performativity

of ethics: Implications for communication governance. *BIG DATA SOC.* **7**, 2053951720915939 (2020).

34. How to research corruption. in *Conference Proceedings Interdisciplinary Corruption Research Forum June* (eds. Schwickerath, A. K., Varraich, A. & Lee Smith, L.) 7–8 (Interdisciplinary Corruption Research Network, 2016).

35. Marzagão, T. Using AI to fight corruption in the Brazilian government. https://files.speakerdeck.com/presentations/9c98a23fd1be410db8b71574a4e852b3/Evidence2Action2017.pdf (2017).

36. Starke, C., Naab, T.K. & Scherer, H. Free to Expose Corruption: The Impact of Media Freedom, Internet Access and Governmental Online Service Delivery on Corruption. *Int. J. Commun. Syst.* **10**, 21 (2016).

37. Harrison, G., Hanson, J., Jacinto, C., Ramirez, J. & Ur, B. An empirical study on the perceived fairness of realistic, imperfect machine learning models. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 392–402 (Association for Computing Machinery, 2020).

38. Kearns, M. & Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. (Oxford University Press, 2019).

39. Rauh, C. Validating a sentiment dictionary for German political language—a workbench note. *J. Inf. Technol. Politics* **15**, 319–343 (2018).

40. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).

41. Taddeo, M. & Floridi, L. How AI can be a force for good. *Science* **361**, 751–752 (2018).

42. Starke, C. & Luenich, M. Artificial Intelligence for EU Decision-Making. Effects on Citizens Perceptions of Input, Throughput and Output Legitimacy. *Data & Policy, 2, E16* (2020).

43. Köbis, N.C., Bonnefon, J.-F. & Rahwan, I. Bad machines corrupt good morals. *Nat Hum Behav* **5**, 679–685 (2021).

44. Acemoglu, D. Harms of AI. *NBER Working Paper* **29247**, (2021).

45. Pinker, S., Nowak, M.A. & Lee, J.J. The logic of indirect speech. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 833–838 (2008).

46. Russell, B. *Power: A New Social Analysis*. (George Allen & Unwin Ltd., 1938).

47. Kalluri, P. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* **583**, 169 (2020).

48. Crawford, K. *Atlas of AI*. (Yale University Press, 2021).

49. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. (Profile Books, 2019).

50. Shortening the risk lifecycle. https://www.microsoft.com/en-us/microsoftacts/shortening-the-risk-lifecycle.

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

52. Kipnis, D. *The Powerholders*. vol. 230 (University of Chicago Press, 1976).

53. Bendahan, S., Zehnder, C., Pralong, F.P. & Antonakis, J. Leader corruption depends on power and testosterone. *Leadersh. Q.* **26**, 101–122 (2015).

54. Keltner, D., Gruenfeld, D.H. & Anderson, C. Power, approach, and inhibition. *Psychol. Rev.* **110**, 265–284 (2003).

55. Bowman, J.S. & West, J.P. Lord Acton and Employment Doctrines: Absolute Power and the Spread of At-Will Employment. *J. Bus. Ethics* **74**, 119–130 (2007).

56. Laskowski, P., Johnson, B., Maillart, T. & Chuang, J. Government Surveillance and Incentives to Abuse Power. *Preprint at https://www.ischool.berkeley.edu/sites/default/files/government-surveillance-abuse-incentives.pdf*.

57. Transparency International. Algorithms in public administration: How do we ensure they serve the common good, not abuses of power? - Blog. *Transparency.org* https://www.transparency.org/en/blog/algorithms-artificial-intelligence-public-administration-transparency-accountability (2021).

58. Mattoni, A. Digital Media in Grassroots Anti-Corruption Mobilizations. in *Sociology and Digital Media* (eds. Rohlinger, D. A. & Sobieraj, S.) (Oxford University Press, 2021).

59. Camaj, L. The media's role in fighting corruption: Media effects on governmental accountability. *Int. J. Press. Polit.* **18**, 21–42 (2013).

60. Köbis, N.C., Iragorri-Carter, D. & Starke, C. A social psychological view on the social norms of corruption. in *Corruption and Norms* (eds. Kubbe, I. & Engelbert, A.) 31–52 (Springer, 2018).

61. Acemoglu, D. & Robinson, J.A. De facto political power and institutional persistence. *Am. Econ. Rev.* **96**, 325–330 (2006).

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

63. Ryman-Tubb, N.F., Krause, P. & Garn, W. How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Eng. Appl. Artif. Intell.* **76**, 130–157 (2018).

64. Crawford, K. *et al.* AI now 2019 report. *New York, NY: AI Now Institute* (2019).

65. Crawford, K. Regulate facial-recognition technology. *Nature* **572**, 565–565 (2019).

66. Oksha, N. Empowering Citizens as Watchdogs in Ukraine - Open Government Partnership. *Open Government Partnership* https://www.opengovpartnership.org/stories/lessons-from-reformers-empowering-citizens-as-watchdogs-in-ukraine/ (2019).

67. Odilla, F. Bots against corruption: Exploring benefits and limitations of AI-based anti-corruption technology. in *International Seminar Artificial Intelligence: Democracy and Social Impacts* (USP Sao Paulo, 2021).

68. Attard, J., Orlandi, F., Scerri, S. & Auer, S. A systematic review of open government data initiatives. *Gov. Inf. Q.* **32**, 399–418 (2015).

69. Mayernik, M.S. Open data: Accountability and transparency. *BIG DATA SOC.* **4**, 2053951717718853 (2017).

70. Flyverbom, M. & Murray, J. Datastructuring—Organizing and curating digital traces into action. *BIG DATA SOC.* **5**, 2053951718799114 (2018).

71. Rafaeli, A., Ashtar, S. & Altman, D. Digital Traces: New Data, Resources, and Tools for Psychological-Science Research. *Curr. Dir. Psychol. Sci.* **28**, 560–566 (2019).

72. Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T. & Oberski, D.L. A framework for digital trace data collection through data donation. *https://arXiv.org/abs/2011.09851v1* (2020).

73. Karakaya, A.-S., Hasenburg, J. & Bermbach, D. SimRa: Using crowdsourcing to identify near miss hotspots in bicycle traffic. *Pervasive Mob. Comput.* **67**, 101197 (2020).

74. Vincent, J. 'Hey Siri, I'm getting pulled over' shortcut makes it easy to record police. *The Verge* https://www.theverge.com/2020/6/17/21293996/siri-iphone-shortcut-pulled-over-police-starts-recording-video (2020).

75. della Porta, D. & Mattoni, A. *Spreading Protest: Social Movements in Times of Crisis*. (ECPR

Press, 2014).

76. Schroth, P.W. & Sharma, P. Transnational law and technology as potential forces against corruption in Africa. *Manag. Decis.* **41**, 296–303 (2003).

77. Murillo, M.J. Evaluating the role of online data availability: The case of economic and institutional transparency in sixteen Latin American nations. *Int. Polit. Sci. Rev.* **36**, 42–59 (2015).

78. Ananny, M. & Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* **20**, 973–989 (2018).

79. Jasanoff, S. Virtual, visible, and actionable: Data assemblages and the sightlines of justice. *BIG DATA SOC.* **4**, 2053951717724477 (2017).

80. Seligsohn, D., Liu, M. & Zhang, B. The sound of one hand clapping: transparency without accountability. *Env. Polit.* **27**, 804–829 (2018).

81. Mittal, M., Wu, W., Rubin, S., Madden, S. & Hartmann, B. Bribecaster: documenting bribes through community participation. in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion* 171–174 (Association for Computing Machinery, 2012).

82. giz. Nigeria: using an app to curb corruption. *Gesellschaft für Internationale Zusammenarbeit (GIZ)* https://www.giz.de/en/mediacenter/39294.html (2016).

83. How Barbie, comedians and new tech are opening up Mexico's government. *Apolitical* https://apolitical.co/solution-articles/en/mexico-labora-supercivicos-open-data.

84. Seering, J., Luria, M., Kaufman, G. & Hammer, J. Beyond Dyadic Interactions: Considering Chatbots as Community Members. in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 1–13 (Association for Computing Machinery, 2019).

85. Grau, P., Naderi, B. & Kim, J. Personalized Motivation-supportive Messages for Increasing Participation in Crowd-civic Systems. in *Proceedings of the ACM on Human-Computer Interaction* vol. 2 1–22 (Association for Computing Machinery, 2018).

86. Li, J., Chen, W.-H., Xu, Q., Shah, N. & Mackey, T. Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology. in *2019 IEEE Global Humanitarian Technology Conference (GHTC)* 1–4 (2019).

87. Savage, S., Monroy-Hernandez, A. & Höllerer, T. Botivist: Calling Volunteers to Action using Online Bots. in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* 813–822 (Association for Computing Machinery, 2016).

88. Cerulus, L. Wikileaks violated privacy rights of hundreds of people: AP. *POLITICO* https://www.politico.eu/article/wikileaks-violated-privacy-rights-of-hundreds-of-people-ap/ (2016).

89. Giles, M. The Cambridge Analytica affair reveals Facebook's "Transparency Paradox". *MIT Technology Review* (2018).

90. Kossow, N. & Miszta, E. *Beyond the hype: Distributed ledger technology in the field of public administration*. https://opus4.kobv.de/opus4-hsog/frontdoor/index/index/docId/3504 (2019).

91. Aggarwal, N. & Floridi, L. The Opportunities and Challenges of Blockchain in the Fight Against Government Corruption. in *19th General Activity Report (2018) of the Council of Europe Group of States Against Corruption (GRECO)* (2018).

92. Bosri, R., Rahman, M.S., Bhuiyan, M.Z.A. & Al Omar, A. Integrating Blockchain With Artificial Intelligence for Privacy-Preserving Recommender Systems. *IEEE Transactions on Network Science and Engineering* **8**, 1009–1018 (2021).

93. AlShamsi, M., Salloum, S.A., Alshurideh, M. & Abdallah, S. Artificial Intelligence and Blockchain for Transparency in Governance. in *Artificial Intelligence for Sustainable*

Development: Theory, Practice and Future Applications (eds. Hassanien, A. E., Bhatnagar, R. & Darwish, A.) 219–230 (Springer International Publishing, 2021).

94. Hyvärinen, H., Risius, M. & Friis, G. A blockchain-based approach towards overcoming financial fraud in public sector services. *Bus. Inf. Syst. Eng.* **59**, 441–456 (2017).

95. Batubara, F.R., Ubacht, J. & Janssen, M. Challenges of blockchain technology adoption for e-government: a systematic literature review. in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* 1–9 (Association for Computing Machinery, 2018).

96. Casino, F., Dasaklis, T.K. & Patsakis, C. A systematic literature review of blockchain-based applications: Current status, classification and open issues. *Telemat. Inform.* **36**, 55–81 (2019).

97. Aarvik, P. *Blockchain as an anti-corruption tool*. https://www.u4.no/publications/are-blockchain-technologies-efficient-in-combatting-corruption (2020).

98. Shang, Q. & Price, A. A blockchain-based land titling project in the republic of Georgia: Rebuilding public trust and lessons for future pilot projects. *Innov. Technol. Gov. Glob.* **12**, 72–78 (2019).

99. Abadi, J. & Brunnermeier, M. Blockchain Economics. *NBER Working Paper* **25407**, (2018).

100. David-Barrett, L. State Capture and Inequality. https://cic.nyu.edu/sites/default/files/cic_pathfinders_state_capture_inequality-2021.pdf (2021).

101. AI procurement in a box. *World Economic Forum* https://www.weforum.org/reports/ai-procurement-in-a-box.

102. Calvaresi, D., Mualla, Y., Najjar, A., Galland, S. & Schumacher, M. Explainable Multi-Agent Systems Through Blockchain Technology. in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* 41–58 (Springer International Publishing, 2019).

103. Tufekci, Z. & Wilson, C. Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. *J. Commun.* **62**, 363–379 (2012).