U4 HELDESK ANSWER 2025:24

Harnessing artificial intelligence (AI) for anti-corruption

Benefits and challenges in prevention, detection and investigation

Miloš Resimić

Reviewed by

Vincent Freigang, Matthew Jenkins, Jon Vrushi and Tom Wright (TI), Daniel Sejerøe Hausenkamph (U4)

Artificial intelligence (AI) is increasingly being applied to anti-corruption efforts, offering significant opportunities and introducing new challenges. Evidence from academic studies and public policy applications shows that AI can strengthen prevention by predicting corruption risks and "hotspots" in public fund allocation, as well as by deterring petty corruption through simplified procedures and citizen access to accurate information. In detection, AI has been used to analyse political integrity datasets and satellite imagery, enabling the identification of risky tenders, fake bidders, conflicts of interest and environmental harms linked to corruption. Though less tested in investigations, AI demonstrates strong potential in evidence gathering and large-scale document review. The most common technologies include machine learning and deep learning, with growing use of natural language processing (NLP) and large language models (LLMs). Key benefits include scalability, predictive capacity and the integration of heterogeneous datasets, while challenges stem from data quality issues, opacity of black-box models, institutional capacity gaps and privacy risks. Ensuring algorithmic transparency, regulatory safeguards and capacity building is essential for responsible deployment.

Helpdesk Answers are tailor-made research briefings compiled in ten working days. The U4 Helpdesk is a free research service run in collaboration with Transparency International.

tihelpdesk@transparency.org

TRANSPARENCY INTERNATIONAL
the global coalition against corruption

Related U4 reading

Artificial Intelligence: A promising anti-corruption tool in development settings (2019)

Artificial intelligence in anti-corruption: A timely update on AI technology (2025)

How to cite

Resimić, M. 2025. Harnessing artificial intelligence (AI) for anti-corruption. Bergen: Transparency International and U4 Anti-Corruption Resource Centre, Chr. Michelsen Institute (U4 Helpdesk Answer 2025)

Published

1 October 2025

Keywords

AI – machine learning – natural language processing – generative AI – corruption prevention – corruption detection – corruption investigation

Query

Please provide an overview of key benefits and challenges of using AI for corruption prevention, detection and investigation.

Main points

- Key features of AI technologies include autonomous learning and task execution, leveraging advanced computing power and the ability to exploit novel data sources like satellite imagery. As such, applying these technologies to anti-corruption efforts can bring benefits such as the reduction of human error, the simplification of repetitive tasks and improved targeting and efficiency in gathering and analysing large datasets.
- Embedding AI – especially LLMs – into existing anti-corruption workflows to automate repetitive data tasks (e.g., converting PDFs to spreadsheets or tabular formats, classifying large datasets, extracting names/dates/entities) delivers immediate value at scale.
- These "mundane" uses of AI in anti-corruption efforts can free up scarce staff time and support activities to prevent (cleaner, richer indicators), detect (faster triage of red flags) and investigate (bulk evidence ingestion and entity extraction) corruption while maintaining a degree of human oversight and control.
- As such, there is a broad scholarly consensus that AI should complement existing anti-corruption workflows rather than replace them. Nonetheless, in addition to the "mundane" uses of AI, there are a growing number of cases of AI tools being developed as a dedicated products to deliver specific outputs in the domain of preventing, detecting and even investigating corruption.
- The potential of AI in corruption prevention has mainly been tested in two areas: (i) predicting corruption risks and "hotspots" in the allocation and disbursement of public funds and (ii) deterring lower-level corruption in public service delivery by simplifying procedures and improving access to accurate information for citizens.
- Key benefits of AI in corruption prevention include scalability compared to traditional audits, meaning that state authorities can allocate their finite oversight resources in a more targeted and efficient manner, which can strengthen safeguards and heighten deterrence. Predictive methods can also help authorities intervene before damage occurs (e.g. before a procurement contract is awarded). These tools can integrate heterogenous datasets and thereby help generate more robust corruption risk profiles.
- AI has also been applied in corruption detection, primarily to strengthen accountability and transparency through political integrity datasets and to use satellite imagery to detect corruption risks linked to environmental harm.
- Key benefits of AI in corruption detection include the ability of AI anti-corruption tools (AI ACTs) to detect high-risk tenders, fake bidders or conflicts of interest among public officials.
- While the potential of AI in corruption investigations has been tested far less, a handful of real-world public policy

applications highlight potentially promising uses in evidence gathering and large-scale document review.

- Across prevention, detection and investigation, academic studies and public policy applications most commonly employ machine learning and deep learning techniques, though natural language processing (NLP) and generative AI – particularly LLMs – have also been applied.

- There are potentially sizeable challenges and drawbacks to the application of AI technologies to anti-corruption efforts. Across the areas of prevention, detection and investigation, key challenges are similar: data quality issues, algorithmic opacity, regulatory lag, lack of institutional capacity, inclusivity gaps, and concerns around privacy and fundamental rights.

- More fundamentally, data limitations and biases can affect the reliability and accuracy of AI ACTs. In one instance from the UK, apparent shortcomings of an AI tool used by the Serious Fraud Office to support its investigations has resulted in past convictions being questioned on procedural grounds, underscoring the importance of rigorously testing the accuracy of AI ACTs before deployment.

- Perhaps most importantly, any shortcomings in training data can compound skewed historical patterns. If corruption is already under-detected in transaction data and AI models are trained only on the limited cases that are identified, the models will become good at detecting those identified positive cases. However, this also reinforces the blind spots, making it harder for the system to detect the many cases that went unnoticed in the first place.

- Moreover, while AI can reduce traditional opportunities for corruption by limiting human discretion, this can also create new risks. If an AI system is flawed or manipulated, it could systematically misclassify cases or misallocate resources at scale before errors are detected. Transparency International (2025) has proposed the concept of "corrupt uses of AI" to cover instances in which AI systems are abused by entrusted powerholders for private gain. This heightens the need for robust oversight, continuous monitoring and transparent algorithmic audit trails to ensure accountability.

- To ensure responsible deployment of AI in anti-corruption efforts, beyond promoting algorithmic transparency and accountability, it is essential to close institutional and regulatory gaps and invest in sustained capacity building initiatives.

Contents

Background

Artificial intelligence (AI) is an umbrella term covering a wide array of technologies and applications, from predictive analytics in finance to autonomous robots in manufacturing, and from medical-image analysis to language translation services. While definitions vary by discipline and use case, most agree on three core features: the ability to sense or interpret an environment; to learn from data; and to act with some degree of autonomy toward specific objectives. For example, the OECD (2024) defines an AI system as:

> "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."

Given AI's many forms and applications, no single, universal definition exists, so this description captures the essential elements where most definitions converge (Jones 2023; Zinnbauer 2025).

The field of AI dates back to the mid-twentieth century, with early work on symbolic reasoning in the 1950s, but its pace of innovation has markedly accelerated in the 2020s with the development of generative AI, thanks to advances in deep learning architectures (e.g. transformers), the explosion of digital data and growing computing power (Roy 2023; Zinnbauer 2025) (see Figure 1). This surge is particularly evident in the development of large foundation models and the technologies that build upon them, such as generative AI.[1] These tools leverage unsupervised or semi-supervised training[2] on vast unlabelled datasets to learn broad patterns before fine-tuning in relation to specific tasks (e.g. GPT-5, LaMDA). This represents a paradigm shift that has unlocked new capabilities in text, image, code and multimedia generation (Roy 2023; Jones 2023).

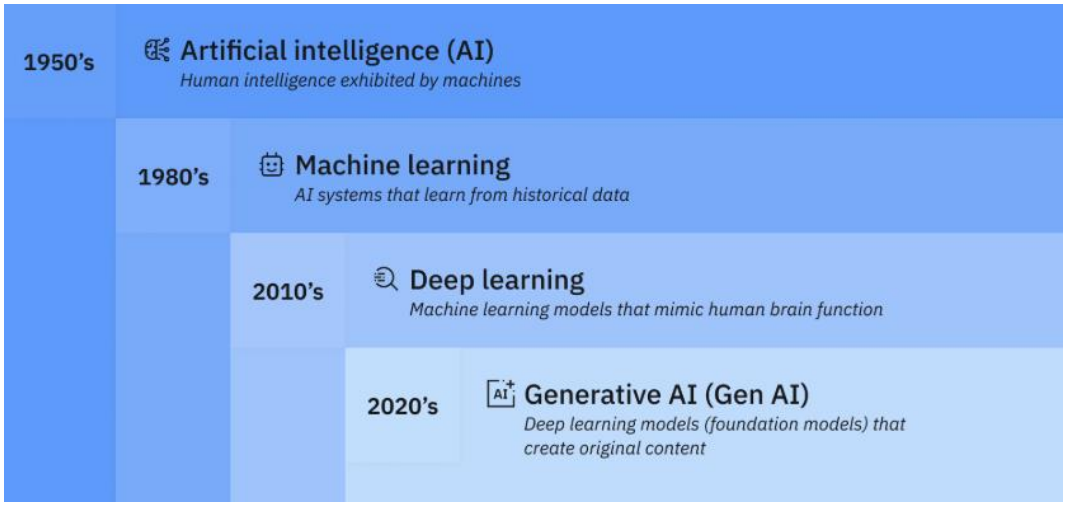
[1] For example, GPT-3.5 and GPT-4 are families of Open AI's GPT foundation models that are trained on a massive amount of text data used to power ChatGPT.

[2] Unsupervised training refers to where an AI model is given only unlabelled data and tasked with discovering hidden patterns or structure in the data without human-provided labels. Semi-supervised training refers to where an AI model is given a small amount of labelled data and a large amount of unlabelled data to improve the model's ability to perform a task, such as classification.

The dramatic growth in both private and public sector support for AI development reflects this momentum. In 2024, generative AI initiatives secured US$33.9 billion in global private investment, an 18.7% increase over 2023, underscoring firms' eagerness to commercialise foundation models and related tools (Maslej et al. 2025).

Governments have likewise ramped up policy and regulation engagement. For instance, US federal agencies issued 59 AI-related regulations in 2024 – more than twice as many as in 2023 – spanning guidance on risk management, data governance and permissible uses of AI in critical sectors (Maslej et al. 2025), though many of these were later repealed in early 2025 by the Trump administration (Wheeler 2025).

Figure 1: The evolution of AI since the 1950s



Source: Stryker and Kavlakoglu 2024

The complex and evolving nature of corrupt practices can make AI-driven solutions appear attractive (Gerli 2024). Research shows that AI has been successful in some traditional anti-corruption areas, including procurement integrity and anti-money laundering, in both preventing and detecting corruption (Zinnbauer 2025). The ability of AI tools to process vast amounts of data, make connections across numerous datasets and learn from data to identify patterns indicative of corruption can aid in the identification of cases for targeted audit as well as the detection of high corruption risk procurement contracts (e.g. López-Iturriaga and Sanz 2018; Mazrekaj et al. 2024; Fazekas et al. 2023; Wacker et al. 2018; García Rodríguez et al. 2022). As such, AI anti-corruption tools (AI ACTs) have been increasingly adopted by governments and private sector organisations around the world. For instance, a bot called Alice has been developed by the Office of the Comptroller General in Brazil to help curb corruption in public procurement (Gerli 2024). For more details on Alice, see page 43 in Annex 1.

However, alongside its promise, AI also introduces significant challenges. First, data reliability and algorithmic bias can arise when models are trained on incomplete, unrepresentative or low-quality datasets, leading to unfair or inaccurate outcomes (Köbis et al. 2022a; Raghavan 2020, Transparency International 2025: 4). Second, inclusivity gaps persist because AI development teams and training data often lack diversity, which can amplify existing social inequities (Leech et al. 2024; Maslej et al. 2025; Transparency International 2025: 4). Third, information asymmetry and opacity in AI decision-making – commonly associated with "black-box" models – undermine transparency and accountability, making it difficult for practitioners and affected communities to understand or challenge AI-driven actions (Odilla 2023, 2024; Transparency International 2025: 4; Zinnbauer 2025). These issues – alongside others discussed in this Helpdesk Answer, such as ethical concerns about surveillance and consent – must be carefully managed if AI ACTs are to achieve their intended impact.

This Helpdesk Answer takes stock of emerging evidence in the academic and policy literature on the benefits and risks of AI ACTs and is structured as follows. The next section explains key AI technologies: machine learning (ML), deep learning, natural language processing (NLP), computer vision (CV) and generative AI (Gen AI). The section after that first introduces how AI can enhance existing anti-corruption work by being embedded into current workflows to augment rather than replace human judgement. The answer then surveys the existing evidence of benefits and risks of using AI ACTs for the prevention, detection and investigation of corruption. The Helpdesk Answer concludes with a discussion of broader challenges for deploying AI in anti-corruption efforts.

AI technologies

Although there are no clear-cut boundaries between different types of AI technologies, they can be generally divided into:

- machine learning (ML)
- deep learning
- natural language processing (NLP)
- computer vision (CV)
- generative AI (gen AI).

Machine learning (ML)

Machine learning (ML) uses algorithms that learn from large amounts of data to identify patterns and make predictions. ML-based models can improve their accuracy over time by autonomously optimising their performance, i.e. they evolve with experience (Syracuse University 2025).

There are many ML techniques or algorithms, chosen according to the problem and the nature of the data (e.g. regressions, decision trees, support vector machines (SVMs), clustering, neural networks) (IBM n.d.). Broadly, ML models fall into three categories:

- supervised learning: human experts label "training" data, which algorithms use to classify or predict outcomes on unlabelled data (IBM n.d.)
- unsupervised learning: does not require labelling because algorithms detect inherent patterns from unlabelled data to cluster the data into groups to inform predictions (IBM n.d.; Murel 2024)
- reinforcement learning: based on trial-and-error learning by systematic rewarding of correct output (IBM n.d.); decision-making is in the hands of autonomous agents which can make decisions in response to the environment without direct instructions from humans (e.g. robots and self-driving cars) (Murel 2024)

ML, along with advanced deep learning techniques, is widely used across domains, including recommendation systems (e.g. Amazon, Netflix), fraud detection in

financial services (e.g. flagging unusual transactions) and personalised marketing (Mate 2025).

In anti-corruption, ML has been shown to be effective in numerous tasks, such as scanning large datasets (e.g. financial transactions, asset declarations and public contracting records) to detect money laundering networks and high corruption-risk areas in procurement processes (Harutyunyan 2023; Gandhi et al. 2024; Katona and Fazekas 2024; Fazekas et al. 2023). Supervised learning[3] has been widely used in anti-corruption research and practice to, for example, detect cartels in public procurement contracting (Fazekas et al. 2023).

Deep learning

Deep learning is a subset of ML that uses multilayered neural networks designed to simulate the complex decision-making processes of the human brain (Stryker and Kavlakoglu 2024; IBM n.d.). These networks consist of an input layer, an output layer and multiple hidden layers of interconnected nodes in between, which allow for the automated extraction of patterns from large, unstructured datasets and the identification of relationships that support predictions (Stryker and Kavlakoglu 2024; Holdsworth and Scapicchio 2024; IBM n.d.).

Deep learning can also be applied in semi-supervised learning, where both labelled and unlabelled data are used to train models for classification tasks (Stryker and Kavlakoglu 2024). Common architectures of neural networks include (Holdsworth and Scapicchio 2024):

- convolutional neural networks (CNNs): designed to detect patterns within images and videos, widely used for pattern and image recognition
- recurrent neural networks (RNNs): typically applied to natural language and speech recognition as they use time-series data to predict outcomes
- generative adversarial networks (GANs): capable of creating new data resembling training data; for example, generated images appearing to be human faces

The most advanced AI applications, such as large language models (LLMs) powering chatbots, rely on deep learning (IBM n.d.). In the anti-corruption field, deep learning

[3] As will be discussed further in this Helpdesk Answer, if corrupt actors learn which indicators trigger high-risk classifications, they can adjust their bidding behaviour and thereby game the system.

has been used, for example, in detecting collusion in procurement using CNNs (Huber and Imhof 2023).

Natural language processing (NLP)

Natural language processing (NLP) is a branch of AI that enables processing, understanding and responding to human language in text or speech form, dealing with large volumes of unstructured textual or spoken data (Syracuse University 2025; Mate 2025). In doing so, it combines ML and deep learning techniques with computational linguistics and rule-based modelling of human language (Stryker and Holdsworth 2024). Not all NLP techniques rely on ML as classical methods like "bag-of-words" are rule-based approaches that do not require training data (Dorash 2017; Stryker and Holdsworth 2024).

NLP typically involves four phases. First, it preprocesses text by transforming it to machine-readable forms (e.g. this can involve techniques such as tokenisation, lowercasing, etc.). Second, the text is transformed into numerical data with which a computer can work using different techniques, such as word embedding. Third, text analysis to interpret and extract meaningful information from text is conducted, including, for example, part-of-speech tagging, aimed at identifying the grammatical roles of words (Stryker and Holdsworth 2024). Finally, processed data is used to train ML models to learn patterns and relationships within the data, which can then be used to make predictions on unseen data (Stryker and Holdsworth 2024).

NLP powers language translation (e.g. Google Translate, DeepL), sentiment analysis (e.g. social media platforms), automated content moderation and virtual assistants (e.g. Siri) (Mate 2025; Dilmegani 2025). It also accelerates mining of information from documents across finance, healthcare, insurance and legal sectors (Stryker and Holdsworth 2024). In anti-corruption, NLP has been applied to analyse large document leaks (e.g. emails, legal documents, whistleblower reports, financial data, corporate ownership). For example, the European Anti-Fraud Office (OLAF) has used NLP techniques to detect suspicious patterns in email correspondence (European Parliament 2021).

Computer vision (CV)

Computer vision (CV) is an AI domain that uses ML and deep learning techniques, especially CNNs, to understand and interpret the content of visual information (i.e. images and videos) (Mate 2025; Syracuse University 2025). By analysing large datasets, CV learns to recognise features and distinguish one image from another (Holdsworth and Scapicchio 2024).

For example, in social media platforms, CV provides suggestions in relation to who might be in a photo posted online. In doing so, ML uses algorithmic models to enable computers to learn about the context of the visual data and learn to distinguish one image from another. A CNN helps ML or deep learning model in this process to "look" by breaking down images into pixels that are labelled to train specific features (IBM 2021; Holdsworth and Scapicchio 2024). The AI model then uses these labels to perform mathematical operations and ultimately predict what it sees and check the accuracy iteratively until the predictions meet expectations (Boesch 2023). The ultimate result is that computers can "see" (e.g. who is in the photo, what objects are on the road) and act based on those insights (Boesch 2023; Holdsworth and Scapicchio 2024).

One example of CV application is the camera lens function of Google Translate that can detect and translate different languages (Holdsworth and Scapicchio 2024). CV is applied in various industries, including energy, utilities, manufacturing and automotive (e.g. real-time object detection, lane following and object tracking in autonomous vehicles) (Holdsworth and Scapicchio 2024).

In the anti-corruption field, CV has been used for satellite data analysis to detect corruption in road construction and mining projects as well as object recognition in public works monitoring (López Acera 2023; Nicaise and Hausenkamph 2025).

Generative AI (gen AI)

Unlike other types of AI technologies, that analyse existing data, gen AI is designed to create new data or content based on a variety of inputs, including text, images, sound, animations and other types of data (Roy 2023). Gen AI models use deep learning neural networks for identifying patterns within the existing data to generate new content (Jones 2023).

Gen AI has broad applications across industries. For example, LLMs are one of the most popular applications of language based generative models that are being used for essay generation, translation and coding, to name a few of their uses (Nvidia n.d.). For instance, Open AI's ChatGPT can be used to draft emails, write essays, code and other tasks, and its abilities keep evolving with the further development of foundation models (Open AI n.d.). For instance, a recent release of GPT-4.5 model reportedly

improves its ability to recognise patterns, make connections and generate creative insights without reasoning, while hallucinating[4] less.

Crucially for anti-corruption, the past one to two years have seen a surge in open-source and open-weight models[5] that can run locally, offering strong task performance without sending sensitive data to external providers (IBM 2023b; AI21 Editorial Team 2025). Examples of open-weight models include Google's Gemma 2, Mistral's Mixtral 8x7B and Alibaba's Qwen 2.5, while notable open-source models include Meta's OpenLLaMA and TII's Falcon Models (Mishra 2025). Local or private deployments enhance data control and confidentiality, allow fine-tuning with domain-specific datasets and can be tailored to narrow tasks (e.g., document classification, entity extraction) that are common in corruption investigations (NetAppInstaclustr 2025).

Gen AI has its potential uses in anti-corruption. For instance, an OECD survey of integrity actors (such as anti-corruption agencies, supreme audit institutions and internal audit bodies) suggests that they see LLMs as especially promising for anti-corruption efforts, particularly with regards to document analysis and text-based pattern recognition (Ugale and Hall 2024). Specifically, the surveyed institutions saw the greatest promise in using LLMs to improve operational efficiency and the analysis of unstructured data. In particular, they highlighted applications in investigations and audits, where anti-corruption bodies must process vast volumes of documents, reports and records (Ugale and Hall 2024). LLMs could assist in evidence gathering and document review by identifying irregularities, extracting relevant information, and flagging suspicious patterns that might otherwise go unnoticed (Ugale and Hall 2024).

[4] Hallucinations in AI refer to instances where a system generates outputs that are incorrect, nonsensical, or entirely fabricated. In large language models (LLMs) and other generative AI tools, this often means producing text that sounds plausible but is factually wrong or unsupported by data. In computer vision, hallucinations may involve perceiving patterns or objects that are not actually present in an image. These errors arise for different reasons including overfitting, bias or inaccuracy of the training data and high model complexity (IBM 2023a). For example, an LLM might confidently cite a non-existent academic article.

[5] Open-weight models are a type of LLM whose parameters are publicly available, allowing anyone to download, inspect, use or fine-tune the model. Unlike fully open-source models, which provide access not only to the weights but also to the model architecture, training code and sometimes the training datasets, open-weight models typically release only the final trained parameters (Mishra 2025).

General benefits and challenges of using AI in anti-corruption efforts

This section first situates AI within the broader evolution of information and communications technologies (ICT) that have been applied to anti-corruption efforts, ranging from early e-governance platforms and electronic procurement systems to today's advanced analytics and predictive tools (Aarvik 2019; Zinnbauer 2025). It then outlines the key benefits and challenges of applying AI tools in anti-corruption contexts.

From digital technologies to contemporary AI systems in anti-corruption

Over the past two decades, the anti-corruption field has increasingly adopted digital technologies to enhance transparency, citizen engagement and oversight (Zinnbauer 2025). Early initiatives focused on e-government systems and the digitalisation of public services, which streamlined administrative processes, reduced waiting times and were empirically shown to lower corruption risks (Shim and Eom 2008; Andersen 2009; Elbahnasawy 2014; Gurin 2014). In parallel, crowdsourcing platforms, whistleblowing portals and open-data dashboards empowered citizens and journalists to monitor procurement, asset declarations and legislative votes in real time (Kossow and Dykes 2018; Kossow and Kukutschka 2017). More recently, blockchain has been piloted to create tamper-proof records of public contracts and financial flows, offering stronger audit trails (Hariyani et al. 2025).

Yet every wave of ICT innovation has brought its own risks. Digital channels can be exploited for illicit transactions on the dark web, crypto-based money laundering or the manipulation of citizen feedback systems (Adam and Fazekas 2018). Today, artificial intelligence represents the next frontier: AI systems can sift through millions of records to flag anomalous bids or suspicious transactions and automate facial recognition or satellite-image monitoring of public works. Early evidence suggests AI can significantly improve detection rates in public contracting and anti-money laundering efforts (Aarvik 2019; Zinnbauer 2025).

However, as with many previous technologies, AI also introduces challenges – such as data quality and bias, algorithmic opacity and uneven access to computing

resources – that must be carefully managed to ensure these tools strengthen, rather than undermine, integrity and accountability.

General benefits of using AI in anti-corruption

Key features of AI systems that can be beneficial to anti-corruption efforts include:

- **Autonomous learning and task execution at scale**: After initial programming and inputs from humans, AI systems can independently train on datasets to execute tasks that once required intensive human effort (Köbis et al. 2022a). Deep learning models using neural networks automatically detect complex patterns and relationships across millions of data points, enabling them to flag subtle corruption indicators that human analysts could miss. For example, research shows that AI ACTs can detect corruption risk zones and predict embezzlement by processing vast volumes of documents and records (Köbis et al. 2022a; López-Iturriaga and Sanz 2018; de Blasio et al. 2022).

- **Leveraging advanced computing power**: modern AI thrives on the exponential growth in computing capacity, which allows it to ingest and process massive datasets in near real time (OECD 2025). This makes it possible to track complex corruption schemes such as the use of shell company networks, aggregate data from disparate sources and run predictive analytics that highlight where bid rigging or embezzlement is most likely to occur (Köbis et al. 2022a; OECD 2025). High-performance computing also enables rich and detailed visualisations of complex financial flows to uncover hidden relationships and money trails (OECD 2025).

- **Fewer human interventions**: unlike human decision-makers, AI, in principle, operates without personal interests, theoretically limiting opportunities for conflicts of interest (Köbis et al. 2022a). By automating decision flows and removing discretionary human steps, AI reduces the number of interactions where petty corruption can occur (Köbis et al. 2022a). However, this reduction in human involvement also means fewer natural audit checks: if an AI system is compromised – through biased training data, for example – there may be no human operator to notice or correct it. In such cases, a corrupted AI could systematically misclassify or misallocate resources at scale before anyone realises something is wrong (Köbis et al. 2022b; Zinnbauer 2025). Thus, while fewer human interventions lower opportunities for traditional corruption, they heighten the importance of robust oversight, continuous monitoring and transparent algorithmic audit trails to detect and respond to any AI system failures or manipulations (Kossow et al. 2021).

- **Capability to integrate novel and disparate data sources**: AI can potentially detect instances of corruption that would be hard to uncover using traditional research methods and tools. For instance, AI systems can leverage large geospatial datasets and satellite imagery to expose corrupt practices in hard-to-reach areas. Deep learning models, for example, have been used to spot illegal road construction linked to deforestation, monitor unauthorised mining and map land grabs by using satellite imagery and deep learning (Labbe 2021; Hausermann et al. 2018; Laurance 2024). These capabilities boost transparency and can empower civil society and local watchdogs with micro-level detail on where illicit activities are occurring, even in remote areas (Zinnbauer 2025; Labbe 2021).

General challenges of using AI in anti-corruption

Key general challenges include:

- **Data reliability and bias**: AI-driven anti-corruption tools are only as good as the data they consume, yet many datasets used in anti-corruption contexts are often incomplete, inconsistent or biased (Köbis et al. 2022a; Zinnbauer 2025). Biases may result from poor quality data (missing values, errors from the data gathering stage) or personal biases of those involved in developing AI systems (Odilla 2024). Both academic studies and public policy AI ACTs analysed in this Helpdesk Answer note the issue of limitations in public expenditure data like public procurement (Fazekas et al. 2023; Katona and Fazekas 2024) and corruption investigation/conviction data (Gallego et al. 2021). These limitations, which include missing data, interoperability challenges, difficult-to-process data formats, as well as deeper concerns like the "selective labelling problem"[6] (see Gallego et al. 2021; Gallego et al. 2022) negatively affect the reliability of AI-driven tools for identifying corruption risks.

  For instance, the project MARA in Brazil, which uses ML to develop individual level corruption scores for civil servants based on previous conviction data, has faced criticism that it simply reinforces existing patterns and biases (Nicaise and Hausenkamph 2025). This is because it was trained only on individuals who were caught and punished, which may exclude undetected corrupt behaviour and focus disproportionately on those civil servants from agencies with robust internal oversight, which consequently overlook broader systemic corruption (Nicaise and Hausenkamph 2025;

[6] The selective labelling problem happens when AI models are trained only on the cases that were caught and labelled (e.g., prosecuted corruption cases), which introduces bias.

Köbis et al. 2022a; Odilla 2024). Data poisoning – deliberate tampering with training data by bad actors – poses another concern, especially for systems that rely on continuously updated online data (Ugale and Hall 2024: 45).

Therefore, it is important to invest in transparency initiatives aimed at ensuring the maintenance of high-quality, unbiased, open government and open data repositories (Zinnbauer 2025: 15). The experience of the Tribunal de Contas (TdC) in Portugal in using AI for developing audit risk models offers several useful lessons on how to improve data quality. First, enhancing data validation mechanisms, which may involve automating the cross-referencing of datasets with other external sources to automatically flag inconsistencies before they can affect the analysis. Second, enforcing data quality and consistency standards across public sector institutions to enhance more efficient use of such data (Hlacs and Wells 2025: 22).

- **Algorithmic challenges and capture**: even the most sophisticated AI models can produce false positives (e.g. wrongly flagging innocent officials) and false negatives (e.g. failing to detect actual corruption). False positives risk reputational harm if AI-generated risk scores or alerts are made public without human verification (Köbis et al. 2022a). Conversely, false negatives can erode trust in AI tools: if a large corruption scandal slips through automated checks, the public may suspect that those in power have tampered with the AI system to hide wrongdoing (Köbis et al. 2022a).

  At the broader level, the weaknesses in the training data can compound blind spots (false negatives). If corruption is already under-detected in transaction data and AI models are trained only on the limited cases that are identified, the models will become good at detecting those identified positive cases. However, this also reinforces the blind spots, making it harder for the system to detect the many cases that went unnoticed in the first place.

  Transparency International (2025) has proposed the concept of "corrupt uses of AI" to cover instances in which AI systems are abused by entrusted powerholders for private gain. This could involve officials trying to game the system. For example, if an official knows what type of behaviour or data would be flagged as a "corruption risk", they could pass on or sell this information to people to help them avoid these supposedly systematic and comprehensive checks. Rather than gaming the system, manipulations can also involve influencing the design of the AI system itself. For instance, algorithmic capture can occur in e-procurement as a one-time manipulation of the AI system by power holders, resulting in narrow politically connected interests reaping long term benefits (Köbis et al. 2022b: 8).

- **Need for human oversight**: effective deployment requires the right balance between AI autonomy and human oversight (Köbis et al. 2022a). Many of the international standards and principles of using AI that have been emerging focus on this aspect. For example, a recently published AI Playbook for the UK Government (2025) calls for a meaningful human control at the right stages, including ensuring that humans validate any high-risk decisions influenced by AI (Principle 4). Governance frameworks – such as the OECD AI Principles and the EU AI Act – also mandate human oversight and thresholds for AI autonomy in high-risk contexts, reinforcing that AI should augment, not replace, human decision-making (OECD 2019; EC 2024).

- **Institutional readiness and regulatory lag**: AI effectiveness in anti-corruption contexts critically depends on the broader institutional, legal and oversight environment in which these AI systems operate (Ubaldi and Zapata 2024). In many jurisdictions, these enabling conditions lag far behind the pace of AI innovation, particularly with the rapid deployment of gen AI since 2022. National and supranational stakeholders have warned that the absence of clear governance frameworks, procurement standards and operational guidelines increases risks of misuse, bias and rights violations (OECD 2019; Ubaldi and Zapata 2024). For example, the EU's AI Act (2024) treats many AI tools as high-risk, thereby obliging AI providers to establish a risk management system, conduct data governance (e.g. ensuring that training, testing and validation datasets are sufficiently representative) and allow for human oversight, among other requests (Future of Life Institute 2024). Yet, in many developing countries, such safeguards are absent or only partially implemented, meaning that AI systems can be deployed without adequate accountability mechanisms in place.

- **Inclusivity gaps**: AI systems' development and training datasets often lack diversity, reinforcing the existing digital divide, which can flow into data, model design and deployment. Women and marginalised communities remain underrepresented among AI professionals, and there is a lack of high-quality training data representing these groups (Zinnbauer 2025). For instance, women are estimated to make up only 29% of AI professionals (Constantino 2024). Moreover, while AI and computer science education is expanding, gaps in access persist as it remains limited in many African countries due to existing infrastructure challenges, including issues as fundamental as access to electricity (Maslej et al. 2025). Considering that most current AI systems are controlled by commercial actors in the Global North (Laforge 2024), they are unlikely to be tailored to the needs of the Global South (Zinnbauer 2025: 15). This increases the risk of discriminatory outcomes and has implications for anti-corruption efforts (Raghavan 2020; Leech et al. 2024; Maslej et al. 2025). For example, when AI tools are trained

on historical "sanctions" or enforcement data, they can reproduce who was already visible to oversight and who was not, thereby amplifying disparities (see Odilla 2024). At the same time, "low-data environments" – common in many countries in the Global South where governance data is patchy – pose additional challenges, since this increases the likelihood that AI ACTs will be trained and rely on datasets with significant gaps or biases (for example, official data that primarily reinforces government narratives) (Foti 2025). This underscores the need to prioritise local data and triangulate information, while strengthening human oversight and ensuring transparency by involving independent experts and civil society actors in AI tool development (Foti 2025).

- **Information asymmetry and opacity**: many AI models, especially deep learning, are "black boxes", producing outputs without transparent explanations (Köbis et al. 2022b: 10). This opacity is compounded by the fact that both the training datasets and the model architectures are often proprietary – created and maintained by private, for-profit vendors – placing them beyond the reach of public scrutiny, although the quality gap between closed and open models have reportedly been shrinking (Pillay 2024). Moreover, this raises acute accountability concerns: if an AI tool generates erroneous or unfair outcomes, it is unclear who should be held accountable: the private developer whose opaque model produced the error or the government body that chose to integrate and act upon those outputs? (see Sanderson et al. 2023).

  On the one hand, the opacity arises from the technical nature of the models, which makes it inherently difficult to interpret results. On the other, it is exacerbated by deliberate secrecy by governments about the AI systems they deploy, as seen with SyRi in the Netherlands and the "Zero Trust" programme in China (Algorithm Watch 2020; Borgesius and van Bekkum 2021; Chen 2019). The challenge of interpretability and explainability[7] is also evident to LLMs, considering the breadth and variety of data fed into these models (Ugale and Hall 2024). Therefore, it becomes challenging to trace the connection between input data and the outputs of these models, making it more difficult for citizens to understand how a decision was made and to appeal those decisions to protect their own rights and interests (Ugale and

[7] When describing interpretability and explainability, it is important to note that while relatively interpretable models such as random forests or regularised regressions can provide feature-level explanations (i.e., highlighting which variables contributed most to a given prediction), these models are highly sensitive to small changes in the dataset. As a result, their "explainability" is limited: they may clarify how the model arrived at a particular prediction, but they tell us much less about the interaction of factors that produced a given outcome.

> Hall 2024: 41). Although there are no easy solutions to these challenges, Ugale and Hall (2024: 41) note that governments have explored the use of decision trees to illustrate the link between the AI systems' results and explanations of how they were reached and have issued explainable AI toolkits, while some academics have introduced a taxonomy of explainable techniques for LLMs (Berryhill et al. 2019; Zhao et al. 2024; Government of the Netherlands 2024).

Without clear governance frameworks, public administrators may be tempted to defer liability to vendors, while developers may evade responsibility by invoking technical complexity. The EU AI Act (2024) and OECD's AI Principles (2019) both emphasise the need for transparency, explainability and shared accountability, requiring that high-risk AI systems include human controls, detailed documentation of training data and mechanisms for contesting decisions (EC 2024; OECD 2019, 2025).

AI and anti-corruption

This section first discusses how AI can enhance existing anti-corruption work by being embedded into current workflows to augment rather than replace human tasks and judgement. In practice, the most common, and immediately valuable, applications are the mundane[8] ones: using AI (especially LLMs) to automate repetitive tasks at scale. The section then reviews empirical evidence on the use of AI in the prevention, detection and investigation of corruption. For prevention and detection, it examines key themes, the AI technologies applied, central findings and the main benefits and challenges, drawing on academic studies and public policy applications. For investigations, where far fewer applications exist, the discussion focuses on the potential benefits and challenges of using AI tools, particularly in evidence gathering and document analysis.

"Mundane" uses of AI technologies in anti-corruption efforts

AI is proving valuable for governments, private firms and non-profits by automating "mundane" tasks that would otherwise consume scarce staff time or be virtually impossible to accomplish due to the vast volumes of data involved. With AI (particularly LLMs), prompts can be adjusted without building complex new programming pipelines, teams can quickly add stages to the existing pipeline and achieve strong accuracy on routine work that would be prohibitively time-consuming – or practically impossible – to perform manually at this scale. Typical examples range from translation, copy editing and project management tasks to converting PDFs into spreadsheets or structured tabular formats, classifying large datasets and extracting names, roles, dates and entities from unstructured text.

As noted later in this Helpdesk Answer, Transparency International UK's (n.d.) use of AI to classify lobbying records transformed vast records of lobbying meetings into an analysable dataset, freeing researchers to focus on substantive insights rather than manual data collection (for practical uses of this dataset see: Whiffen 2025).

Similarly, the Transparency International Global Health Atlas aggregates and structures diverse information on corruption in the health sector, creating a reusable

[8] This "mundane" layer is cross-cutting: it underpins prevention (cleaner, richer indicators), detection (faster triage of red flags) and investigation (bulk evidence ingestion and entity extraction), enabling experts to examine far more material with the same resources.

evidence base for investigations, advocacy and policy design (Transparency International Global Health 2025). These initiatives have demonstrated the usefulness of LLMs, ML and NLP based techniques to turn vast amounts of information into structured and regularly updated datasets.

Box 1. Civil society uses of AI in anti-corruption efforts: The experience of Transparency International UK

For several years, Transparency International UK has moved beyond theoretical work to deploy practical AI tools that turn vast, complex datasets into actionable evidence addressing public integrity challenges, from lobbying and conflicts of interest to public procurement corruption. In doing so, they have applied a range of AI techniques, including ML, NLP and gen AI.

Their Health Atlas project (discussed in the following section) uses LLMs in a multi-stage classification process. The system identifies articles relevant to corruption, classifies the specific type of corruption (e.g., bribery), and extracts key metadata like location and date, creating the world's largest repository of evidence on health corruption. In the project's early stages, they also analysed the data (using BERT sentiment analysis[9]) to distinguish between reports of corruption and articles detailing anti-corruption efforts.

Building on these capabilities, TI UK developed Tomni AI, a versatile platform that integrates retrieval-augmented generation. This allows users to ask natural language questions of vast document sets. This system uses a hybrid approach, combining vector-based semantic search with traditional keyword search (BM25) for accuracy, before an LLM synthesises the retrieved information into a coherent answer. Tomni AI is also designed as an agentic AI system, where a researcher can define a customed sequence of tasks – such as semantic chunking, data enrichment with user-defined red flags (e.g. extract when suspicious transactions are noted and their value), and named entity and relationship extraction (to automatically populate network maps from vast PDFs/leaks/investigations) – for the system to execute automatically on unstructured data.

[9] BERT sentiment analysis uses a model called BERT (Bidirectional Encoder Representations from Transformers) to understand and classify the sentiment of text, such as determining if a review is positive or negative. It leverages deep learning techniques to analyse the context of words in sentences, improving the accuracy of sentiment detection.

AI and prevention[10] of corruption

Key themes

The potential of AI technologies in preventing corruption has been tested primarily in two broad thematic areas:

- predicting corruption risks and corruption "hotspots" in the allocation and disbursement of public funds (Huber and Imhof 2023; Gallego et al. 2021; Decarolis and Giorgiantonio 2022; Rodríguez et al. 2022; Fazekas et al. 2023; Katona and Fazekas 2024; Mazrekaj et al. 2024)

- deterring lower-level corruption in the public sector (Odilla 2023; Köbis et al. 2022a; TNRC 2024).

Academic studies and public policy applications of AI ACTs demonstrate the potential of AI to identify corruption risks in public procurement contracting (e.g. Gallego et al. 2021; Odilla 2023; Katona and Fazekas 2024). Research also shows AI can effectively detect cartels and collusion in public contracting (e.g. Huber and Imhof 2023; Fazekas et al. 2023). Furthermore, empirical studies and public policy deployments highlight AI's potential to uncover conflicts of interest, either by identifying politically connected firms (Mazrekaj et al. 2024) or by classifying lobbying meetings and health corruption cases into various categories (Transparency International Global Health 2025; Transparency International UK n.d.).

AI has also been applied to simplify administrative procedures and improve public awareness of public service delivery, thereby reducing opportunities for petty corruption. A notable example is Justina del Mar, an NLP based chatbot developed by WWF Peru. It provides artisanal fishers and shipowners with instant, reliable answers about inspections, licensing, sanctions and reporting channels (WWF 2024). Built from phone surveys and consultations that mapped common questions and knowledge gaps then systematised with regulatory texts and information obtained from authorities, the tool provides plain-language guidance via simple menus or natural language queries (WWF 2024). By putting accurate rules and reporting channels directly into users' hands, it is intended to reduce opportunities for petty bribery or extortion tied to information asymmetries and to strengthen transparency and trust between communities and authorities (WWF 2024).

[10] In this Helpdesk Answer, prevention involves those cases in which AI can identify corruption "hotspots", whereas detection covers the cases where AI can identify specific entities (i.e. public officeholders, firms, public procurement contracts).

AI technologies in corruption prevention

Most academic studies and public policy applications rely on machine learning (ML) techniques, ranging from traditional regressions to deep learning models (e.g. Fazekas et al. 2023; Gallego et al. 2022; García Rodríguez et al. 2022; Mazrekaj et al. 2024). Existing academic studies often employ linear models, such as lasso, and tree-based ensemble models, like random forests, while fewer use neural networks (e.g. López-Iturriaga and Pastor Sanz 2018).

Some initiatives also use natural language processing (NLP) techniques, which were shown to be especially useful in generating novel corruption risk indicators based on textual data from public procurement notices (e.g. Katona and Fazekas 2024). Moreover, Transparency International UK has been using gen AI (LLMs) for their projects on classifying lobbying meetings (Annex 1, Example #13) and healthcare related corruption stories (Annex 1, Example #12).

Across approaches, trade-offs emerge: interpretable models (e.g. traditional statistical models) facilitate transparency and policy adoption, while more complex black-box models (e.g. neural networks) can increase predictive performance but risk resistance from oversight bodies due to their opacity.[11]

Central findings

Evidence from academic studies and practical deployments analysed in this Helpdesk Answer suggests that AI ACTs show promising results in preventing corruption. Most academic studies report that their models achieve high predictive performance in identifying corruption "hotspots" in public procurement, detecting collusive bidding patterns and flagging politically connected firms with the potential of predicting conflicts of interest (e.g. Katona and Fazekas 2024; Fazekas et al. 2023; Mazrekaj et al. 2024; Huber and Imhof 2023).

For instance, using NLP and ML algorithms to analyse the text in tender documents lifted single-bid prediction[12] from 77% to 82% in Hungary (Katona and Fazekas 2023), while ML models correctly predict over 85% of politically connected firms in the Czech Republic based solely on firm-level financial and industry indicators (Mazrekaj et al. 2024).

[11] There are also models that aim to combine interpretability with strong predictive performance. One notable example is the explainable boosting machine, a glass-box model with built-in interpretability: it is transparent by design, enabling users to understand its logic, while often achieving accuracy comparable to state-of-the-art black-box models (InterpretML n.d.; McGrail 2023; MinnaLearn n.d.).

[12] The number of tenders that received only one bid.

Key benefits and challenges

The key advantage of AI technologies for corruption prevention lies in their superior efficiency and scalability compared to traditional audits. By prioritising corruption risk "hotspots" for targeted audits, authorities can allocate their finite oversight resources more efficiently (Katona and Fazekas 2024; Fazekas et al. 2023). Complex models – such as gradient boosting or neural networks – can uncover subtle, non-linear patterns in procurement or financial data that simple rule-based systems or human auditors may miss (Chen 2019; Gallego et al. 2021). Predictive scoring also enables authorities to intervene before contracts are awarded, reducing losses from corrupt practices, such as in the case of the Alice tool in Brazil (Odilla 2023). Moreover, AI tools can integrate heterogeneous data sources – structured procurement records, corporate ownership registries, social network mappings and unstructured text – to create richer and more robust corruption risk profiles and identify more complex collusion and conflicts of interest patterns (Aarvik 2019; García Rodríguez et al. 2022; Mazrekaj et al. 2024; Zinnbauer 2025).

However, challenges persist. First, the opacity of black-box models such as deep neural networks limits interpretability, which can undermine public trust (Chen 2019). Second, there are still significant challenges with data quality, especially with regards to public procurement datasets (Katona and Fazekas 2024; Fazekas et al. 2023). Incomplete, inconsistent or biased datasets can produce misleading results, amplifying false positives or negatives. This is especially problematic with the use of corruption investigation/conviction data, which is prone to a "selective labelling problem" (see Annex 1, Example #9 below) (Gallego et al. 2021). Third, corrupt actors may adapt their behaviour once risk indicators become known, requiring periodic retraining with updated data to maintain model effectiveness (Fazekas et al. 2023). Finally, there are resource and capacity gaps: building and sustaining these systems requires technical expertise and institutional infrastructure that may be lacking in some public agencies. In citizen-facing tools, such as Justina del Mar in Peru, basic digital literacy among end-users is also a prerequisite for effective use (TNRC 2024).

For a list of examples of the application of AI ACTs in corruption prevention, see Annex 1.

AI and detection of corruption

Key themes

The potential of AI technologies in detecting corruption has been tested in two broad thematic areas:

- strengthening public sector accountability and transparency by leveraging and connecting diverse political integrity datasets (Odilla 2023; Harutyunyan 2023; Chen 2019; Transparency International Ukraine 2025; Bosisio et al. 2021; Wacker et al. 2018; Gandhi et al. 2024)
- using satellite imagery to detect corruption risks linked to environmental harm (WWF 2023; Wageningen 2023; Hillsdon 2024; Labbe 2021; Paolo et al. 2024; Zinnbauer 2025)

In the first area, academic studies and public policy applications demonstrate AI's potential to detect corruption risks associated with firms and individual public officeholders. Tools focused on individuals have scrutinised asset declarations (Harutyunyan 2023), flagged wasteful spending (Odilla 2023) and detected conflicts of interest or irregular financial movements (Chen 2019). Firm-focused applications have assessed corruption risks in public procurement contracts (Transparency International Ukraine 2025), flagged corporate ownership anomalies (Bosisio et al. 2021), detected fake suppliers (Wacker et al. 2018) and identified money laundering patterns (Gandhi et al. 2024).

For instance, the Tweetbot Rosie de Serenata was created to analyse reimbursement claims submitted by members of Brazil's congress (Köbis et al. 2022a). Originating from the grassroots anti-corruption initiative Operação Serenata de Amor (Operation Love Serenade), Rosie processes official spending data and flags suspicious expenses, such as cases where a congress member appears to have been in two locations on the same day and at the same time (Cordova and Gonçalves 2019). When a suspicious transaction is detected, Rosie automatically posts the finding on X, inviting citizens and legislators to confirm or refute the suspicion (Cordova and Gonçalves 2019).

In the second area, AI combined with satellite imagery has been used to detect illegal deforestation (WWF 2023; Wageningen 2023; Hillsdon 2024), illegal mining (Labbe 2021) and illicit fishing activities (Paolo et al. 2024), among its other uses (see Zinnbauer 2025). For instance, Forest Foresight, developed by WWF-Netherlands with commercial and academic partners, has demonstrated the ability to predict illegal deforestation up to six months in advance with 80% accuracy in pilot projects in Gabon and Borneo (WWF 2022).

AI technologies

Most public policy applications of AI ACTs and academic studies analysed in this Helpdesk Answer targeting corruption risks among public officeholders and firms use ML algorithms (Odilla 2023; Chen 2019; Transparency International Ukraine 2025). In contrast, applications based on satellite imagery typically use deep learning neural networks (CNNs) for image classification (Wacker et al. 2018; Paolo et al. 2024). The former typically favour relatively interpretable models (e.g., random forests and regularised linear/logistic models) that allow feature-level explanations and audit trails, whereas CNN based image classifiers are higher-accuracy but comparatively more opaque.

Some public policy applications combine NLP techniques with ML and deep learning neural networks, for example, for text pattern recognition. An example is Inspector AI in Peru, which uses NLP to process suspicious transaction reports (GIZ 2024; European Parliament 2021; Gerli 2024; Nicaise and Hausenkamph 2025).

Central findings

Evidence from public policy applications of AI ACTs and academic studies show promising results in AI's ability to detect corruption, although some public policy applications are still in their early stages or are currently undergoing further development.

Civic-tech platforms such as Rosie (Brazil) (Odilla 2023; Operação Serenata de Amor n.d.), and DOZORRO (Ukraine) (see Box 2) have helped prevent inefficient spending, blocked corruption-prone contracts before signature and uncovered conflicts of interest among public officials. AI ACTs have also enabled civil servants to detect corruption risks faster by automating processes, such as extracting information from suspicious transaction reports in Peru (GIZ 2024) and automating the process of early detection of illegal deforestation (e.g. WWF 2023; WWF Ecuador 2025). Other deployments, like China's Zero Trust programme, have demonstrated large-scale anomaly detection capacity by cross-referencing extensive government datasets, although this has raised privacy concerns (Odilla 2023; Chen 2019; Transparency International Ukraine 2025).

Academic studies further show AI's potential to detect corruption, for example, by detecting potentially illegal fishing activity by mapping industrial fishing vessels that are not captured by public monitoring systems (Paolo et al. 2024) or by detecting fake suppliers in public contracting (Wacker et al. 2018).

Box 2. Countering public procurement corruption with DOZORRO in Ukraine

DOZORRO is a feedback platform integrated with Ukraine's centralised public procurement database, Prozorro, established in 2016, which uses machine learning to monitor public procurement activities (Strawinska n.d.; Kucherenko 2019). It is an AI-powered system of civic monitoring over public procurement in Ukraine developed by Transparency International Ukraine (2018). Unlike systems that rely on exhaustive lists of corruption risk indicators – an approach vulnerable to gaming of indicators or manipulation by corrupt officials – DOZORRO's software was designed to avoid such predictability (Transparency International Ukraine 2018).

In 2018, 20 experts reviewed approximately 3,500 tenders without access to the amounts or names of procuring entities, simply indicating whether each tender appeared risky. Their assessments were then used to train the AI algorithm (Transparency International Ukraine 2018). Today, the system independently evaluates the likelihood of corruption risks in tenders and forwards high-risk cases to civil society organisations within the DOZORRO network. The model "remembers" correct classifications and "forgets" incorrect ones, enabling continuous refinement over time (Transparency International Ukraine 2018; Aarvik 2019).

DOZORRO has proven effective in preventing inefficient spending and identifying high corruption risk contracts in Ukraine (Open Stories 2021; Transparency International Ukraine 2025a).For instance, in July 2025, the DOZORRO team analysed 165 procurement procedures worth over UAH 10 billion (approx. €207 million) and identified violations in 89 tenders, with inflated pricing the most common issue (Transparency International Ukraine 2025b). Through DOZORRO's requests to procuring entities and its cooperation with law enforcement in the first half of 2025, inefficient spending of UAH 133 million (approx. €2.7 million) was prevented (Transparency International Ukraine 2025b).

Key benefits and challenges

The key benefit of deploying AI technologies for corruption detection is the ability of AI ACTs to detect risky tenders (Transparency International Ukraine 2025), fake bidders (Wacker et al. 2018), suspicious financial transactions that could indicate money laundering (Gandhi et al. 2024) or conflicts of interest of public officials (Chen 2019; Harutyunyan 2023) before money moves. In addition, these tools can detect environmental challenges such as illegal deforestation (WWF 2023; WWF Ecuador 2025) before environmental degradation occurs, enabling early audits and inspections.

These tools use NLP techniques, ML and deep learning algorithms to sift through and connect heterogenous datasets at a high scale and speed to look for asset declaration discrepancies (Harutyunyan 2023), procurement suppliers' corruption risks (Hlacs and Wells 2025), financial transaction data to identify money laundering patterns (Gandhi et al. 2024) or satellite data to identify potential illegal fishing (Paolo et al. 2024), mining (Labbe 2021) or deforestation (WWF 2023). This can free up human resources to focus on high-risk cases.

Documented impact of these tools includes reduced meal reimbursement spending by approximately 10% after the launch of Rosie in Brazil (Odilla 2023; Operação Serenata de Amor n.d.), twice as many cases referred to prosecutors with Peru's Inspector AI (GIZ 2024) and ranger interventions guided by Forest Foresight risk maps (WWF 2023). This is evidence that AI can translate into concrete enforcement outcomes.

Challenges are, however, significant. As with prevention, data quality remains a central obstacle: incomplete, inconsistent or biased datasets undermine reliability (European Parliament 2021; Gandhi et al. 2024; Paolo et al. 2024; Hlacs and Wells 2025).

Further, the use of high-performance but less-explainable models (e.g. based on neural networks like CNNs) is even more present in detection compared to prevention (see examples below), which carries a risk of public backlash or institutional resistance (Chen 2019). The stakes are also significantly higher in detection: unlike prevention models, which often highlight broad risk patterns or "hotspots," detection tools focus on specific individuals, firms or entities. As a result, errors can have grave consequences: false positives may wrongly implicate officials or companies, damaging reputations, prompting lawsuits or even leading to unjust legal action, while false negatives can allow serious corruption to go undetected. This combination of opacity and high-stakes outcomes makes explainability, transparency and careful human oversight particularly critical at the detection stage.

There are also operational and capacity challenges. The benefits of AI tools depend on embedding them into workflows (case management, audit triggers), sustained funding and skills to retrain/monitor models, plus on-the-ground capacity to act on alerts (e.g., deforestation hotspots) (WWF 2023).

Corrupt actors may also adapt their behaviour once risk indicators become known, requiring regular model retraining against evolving corrupt patterns (see Fazekas et al. 2023).

Finally, some corruption detection tools raised serious privacy concerns, such as SyRi in the Netherlands (Algorithm Watch 2020; Borgesius and van Bekkum 2021) and the Zero Trust programme in China (Chen 2019), due to secrecy of the algorithm used, a lack of transparency on how private data is handled and linkage of vast data sources.

For a list of examples of the application of AI ACTs in corruption detection, see Annex 2.

AI and corruption investigations: Potential, benefits and challenges

Compared to prevention and detection, investigative AI based tools are less common. According to some recent surveys (Ugale and Hall 2024), real-world applications are still limited, particularly in the public sector, and the return on investment remains unclear.

Nonetheless, the potential of AI technologies in corruption investigations is increasingly recognised. For instance, in its 2024 report, the European Public Prosecutor's Office highlights that its digital operations team initiated the Operational Digital Infrastructure Network programme to develop digital tools that would strengthen their investigators' capacities by using artificial intelligence and big data analysis (EPPO 2025). Similarly, the European Investment Bank's investigations division engaged with international investigators in 2024 to explore the use of AI in investigations (EIB 2025).

A recent OECD survey of 59 organisations across 39 countries (Ugale and Hall 2024) found strong interest in generative AI (particularly LLMs) for anti-corruption work. Respondents highlighted their potential to make auditors' and investigators' work more efficient by automating time-consuming tasks and allowing staff to focus on activities requiring human judgement and expertise (Ugale and Hall 2024: 15). Surveyed integrity actors identified the greatest value of LLMs in the areas of investigative and audit processes as relating to evidence gathering and document review (Ugale and Hall 2024: 17). Specifically, LLMs could help investigative work by organising large volumes of text for easy prioritisation and consumption and help in pattern recognition. Indeed, several AI tools have already been deployed in major corruption investigations (see Box 3).

Box 3. AI in corruption investigations: from Operation Car Wash to Rolls Royce bribery cases

In Brazil, the ContÁgil system, a data retrieving and data analysis tool for identifying fiscal fraud and money laundering, developed by the Special Secretariat of Federal Revenue of Brazil (RFB) was used in one of the largest corruption investigations in Brazil, the Operation Car Wash (Lava Jato), to identify complex networks of intermediaries and shell companies and connecting them with politicians and businessmen (Jambeiro Filho 2019; Odilla 2023: 378). It uses data such as asset

declarations, ownership registers and tax payments, among other sources (Odilla 2023: 363).

Its resources include an ML environment with supervised learning algorithms and algorithms for clustering, outlier detection, topic discovery and co-reference resolution (Jambeiro Filho 2019). ContÁgil also has a social network analysis tool, which can visualise networks between people and companies based on scanning various data sources (Jambeiro Filho 2019).

The tool increases efficiency dramatically as it reportedly accomplishes in one hour what would take a human inspector a whole week (Jambeiro Filho 2019). The tool however, also exhibited some challenges, such as alert overflow, false positives, slow adaptability to new forms of wrongdoings and limited auditability, among other challenges (Odilla 2023: 372).

AI has also been used by the Serious Fraud Office (SFO) in the UK. In 2018, the SFO hired an "AI lawyer" tasked with automatically analysing documents (Martin 2018). The SFO previously used similar technology developed by a Canadian firm to spot legally privileged information among 30 million documents during the four-year long Rolls Royce bribery and fraud investigation (Martin 2018; Royas 2017; López Acera 2023). The SFO noted that the technology was 80% cheaper than hiring outside counsel to review documents and identify legally privileged information (Martin 2018).

The OpenText Axcelerate, an AI-driven tool, not only flags legally privileged materials but scans and organises information from various data formats and document types, displaying relevant information for investigation (Martin 2018). However, a recent report highlighted several challenges associated with this and similar tools. In February 2025, SFO noticed that searches did not return expected results as number of documents were omitted due to formatting issues, which required reconfiguration of the software, demonstrating the need for regular maintenance of AI tools (Fisher 2025: 75; Herbert Smith Freehills Kramer n.d.; Ring 2025).

A more serious concern relates to lawyers' recently questioning of SFO's evidence disclosure software, which could undermine historical convictions (Ring 2024). Specifically, SFO is investigating its legacy software tools (its old provider Autonomy Introspect and current Axcelerate system) that were used on dozens of cases to evaluate how it recognised punctuation, and it is reviewing the encoding issue which may have interfered with document searches (Ring 2024). Lawyers have been asking for clarity from prosecutors about how these issues may have affected current and past cases in terms of disclosure (Ring 2024).

While integrity actors considered document analysis and text-based pattern recognition the most valuable uses of LLMs for anti-corruption and anti-fraud investigations, they reported few advanced initiatives (Ugale and Hall 2024: 7). According to the survey, the actors who use LLMs either rely on turnkey foundation models[13] developed by private firms or adapt existing models by fine-tuning them with specific datasets for targeted tasks (Ugale and Hall 2024: 8).

[13] A model which is available in a ready to use form, without a need for fine tuning (Ugale and Hall 2024: 32).

Broader challenges and implications of AI in anti-corruption efforts

Addressing risks of potentially harmful uses of AI systems

As highlighted in previous sections, including the Zero Trust programme in China, many AI ACTs involve centralised government control over sensitive data. This raises significant risks of surveillance, political capture and consolidation of power by political officeholders and tech companies at the expense of public interest (Köbis et al. 2022b; Köbis 2023). Similar concerns have been raised about the consolidation of power in the hands of small number of highly influential tech companies developing AI systems, who "are likely to steer the technology in a direction that serves their narrow economic interests rather than the public interest" (von Thun 2023).

In such contexts, AI systems may even be deliberately abused by power holders for private gain at the expense of the public interest through the intentional design of AI systems for corrupt purposes, manipulation of training data or corrupt application of otherwise legitimate tools (Köbis et al. 2022b: 7). Transparency International (2025: 3) has sought to draw attention to these risks through its advancement of the concept of "corrupt uses of AI", defined as the abuse of AI systems by entrusted powerholders for private gain. Commercial vendors can also exacerbate these risks by promoting opaque, top-down solutions that create vendor lock-in,[14] inefficiencies or undue influence (Köbis et al. 2022a).

A first line of defence lies in ensuring algorithmic transparency[15] and accountability, recognising that algorithmic systems are far from being neutral forms of technology (Zerilli et al. 2019; Jenkins 2021). Kossow et al. (2021: 11) highlight two key goals of algorithmic transparency and accountability: understanding how an algorithmic system generally functions and clarifying how it reaches individual outcomes. To this

[14] Referring to a situation in which a customer is dependent on a single vendor for a product or service, making it difficult and costly to switch to an alternative provider due to proprietary technologies.
[15] As Kossow et al. (2021: 11) note, there is no precise definition of what constitutes algorithmic transparency as these systems can come with different degrees of transparency depending on their technical properties and governance processes.

end, the Association for Computing Machinery (2017: 2) proposes seven key principles for algorithmic transparency and accountability:

1. "Awareness: owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.
2. Access and redress: regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.
3. Accountability: institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.
4. Explanation: systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts
5. Data provenance: a description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorised individuals.
6. Auditability: models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.
7. Validation and testing: institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public."

Complementing these principles, emerging integrity frameworks for public administration emphasise that (see Jenkins 2021; Kossow et al. 2021):

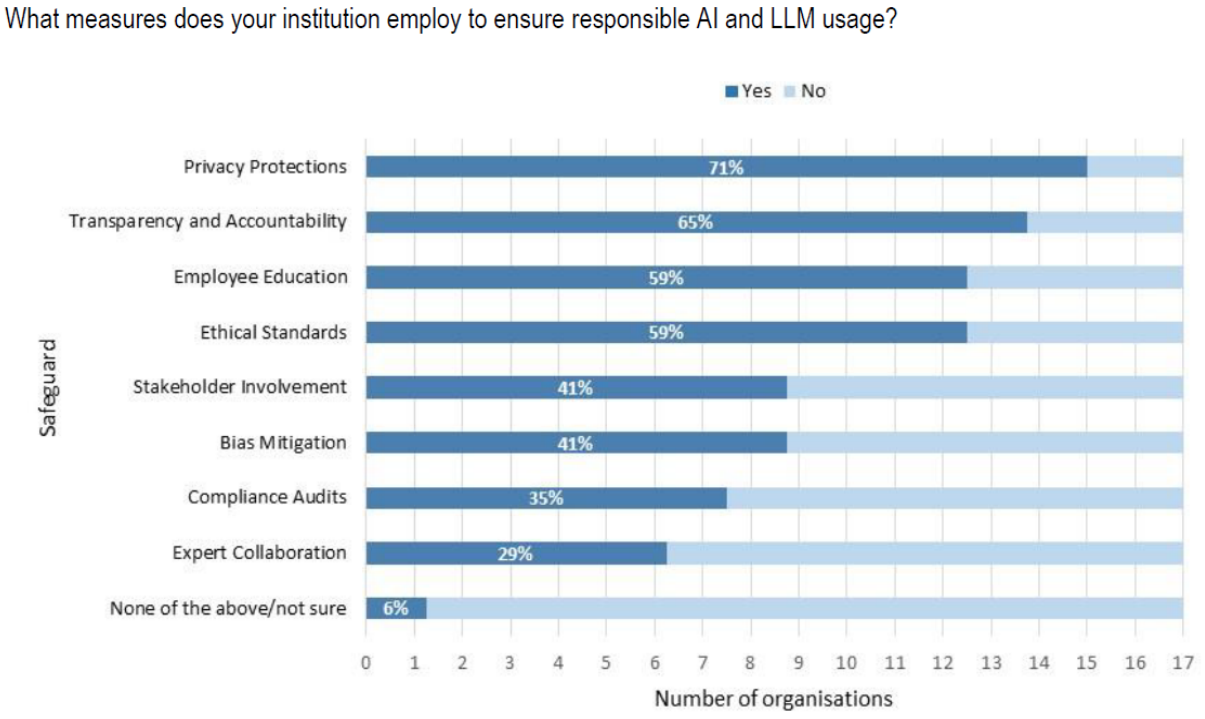
- people have to be informed when they interact with or are subject to an AI system
- individuals affected by an AI decision must be helped to understand its outcome

- those who are affected by an AI system decision should be able to challenge the outcome

Köbis et al. (2022b) further stress the importance of data and code transparency, routine model audits, ethics training for data scientists and strengthened whistleblowing mechanisms to minimise the risk that AI tools are misused or even used themselves to perpetrate corrupt acts.

There is some limited emerging evidence of responsible AI practices beginning to be deployed in certain countries. An OECD survey of integrity actors (such as anti-corruption agencies, supreme audit institutions and internal audit bodies) on their use of gen AI found that most of the institutions surveyed already employ privacy protections, transparency and accountability measures and employee education as central safeguards for ethical use of AI tools and LLMs (Ugale and Hall 2024: 35) (see Figure 2).

Figure 2. Reported safeguards for ensuring responsible use of AI and LLMs.



Note: Possible responses included the following: 1) Ethical Standards: Implementation of ethical guidelines and policies; 2) Transparency and Accountability: Ensuring open AI decision making and maintaining accountability; 3) Bias Mitigation: Actively addressing biases to promote fairness; 4) Privacy Protections: Adhering to privacy and data protection standards; 5) Compliance Audits: Conducting regular ethical and legal compliance assessments; 6) Stakeholder Involvement: Engaging with stakeholders for input and addressing concerns; 7) Employee Education: Offering training and awareness programmes on responsible AI; 8) Expert Collaboration: Working with external experts for ethical and legal guidance; 9) None of the above/not sure; 10) Other. None of the respondents selected other.
Source: OECD questionnaire

Source: Ugale and Hall 2024: 35.

As Gerli (2024) notes, there is currently a broad scholarly consensus that AI should complement existing anti-corruption efforts rather than replace them, and that the overall ecosystem in which AI-driven solutions are designed, developed and implemented is critical for its success, as will be discussed in the following section (Etzioni and Etzioni 2017; Köbis et al. 2022a).

Addressing institutional and regulatory gaps

Without a broader institutional framework, it is difficult to ensure that AI systems for anti-corruption are procured, designed and used in a responsible manner (Gerli 2024). Over the past five years in particular, new standards and pieces of legislation specifically addressing AI have emerged (see Box 4).

Box 4. Standards and principles for the development and deployment of AI systems in public administration

There is a growing body of international standards aimed at ensuring the responsible use of AI in public administration, with a focus on accountability and public integrity (Ubaldi and Zapata 2024; OECD 2025). In parallel, an increasing number of countries is adopting national level regulations and principles (Ugale and Hall 2024; Maslej et al. 2025; UK Government 2025). These frameworks seek to mitigate risks of bias, opacity, misuse and lack of accountability by embedding transparency, human oversight, explainability and rights protection into AI systems (OECD 2019).

The OECD Principles on Artificial intelligence (2019, updated in May 2024) were the first intergovernmental standards on AI, setting out five core principles for policy makers and AI actors: 1) inclusive growth, sustainable development and well-being; 2) human rights and democratic values, including fairness and privacy; 3) transparency and explainability; 4) robustness, security and safety; and 5) accountability. For example, the transparency and explainability principles require AI actors to provide meaningful information about AI systems – their data sources, logic, and limitations – so that those affected can understand and challenge automated decisions (OECD 2019). Accountability further requires traceability of datasets, processes and decisions (OECD 2019).

The EU AI Act (2024) is the world's first comprehensive framework for AI regulation (European Parliament 2023; Elbashir 2024). It establishes obligations for AI systems based on risk levels, distinguishing between four risk levels (Ubaldi and Zapata 2024: 16):

- unacceptable risk: prohibited uses (e.g. predictive policing, social scoring or assessing the risk of an individual committing criminal offences)

- high-risk: uses with significant potential harm to health, safety, democracy (e.g. most public sector applications), which therefore require establishing a risk management system, data governance, compliance documentation and fundamental rights impact assessments, among other requirements)
- limited risk: applications such as chatbots and deep fakes, where transparency obligations require informing users that they are interacting with AI
- minimal risk: systems which require code of conduct, such as video games.

These measures prioritise transparency obligations to ensure AI systems' accountability and explainability (Elbashir 2024).

At the national level, several initiatives complement these frameworks. In 2025, the UK government published its artificial intelligence playbook, which sets out ten principles for safe, responsible and effective use of AI in government organisations. In January 2024, the Netherlands became one of the first countries to publish a strategy specifically focused on gen AI, outlining a vision for using gen AI in the public sector (Ugale and Hall 2024: 14). These principles include developing and applying gen AI in a safe way, developing and applying it equitably, ensuring that it serves human welfare and safeguards human autonomy, and that it contributes to sustainability and prosperity (Ugale and Hall 2024: 14).

Other countries have advanced oversight mechanisms for the deployment and use of AI systems in the public sector: Norway's office of the auditor general began auditing central government AI use in 2023; France published a guide on algorithmic transparency; Ireland issued guidelines for AI in the public service; and the Netherlands developed governance guidance for responsible AI applications (Ubaldi and Zapata 2024: 15, 19).

Procurement of AI systems also requires clear standards. A recent Transparency International (2025: 10) paper addresses corrupt uses of AI and recommends that:

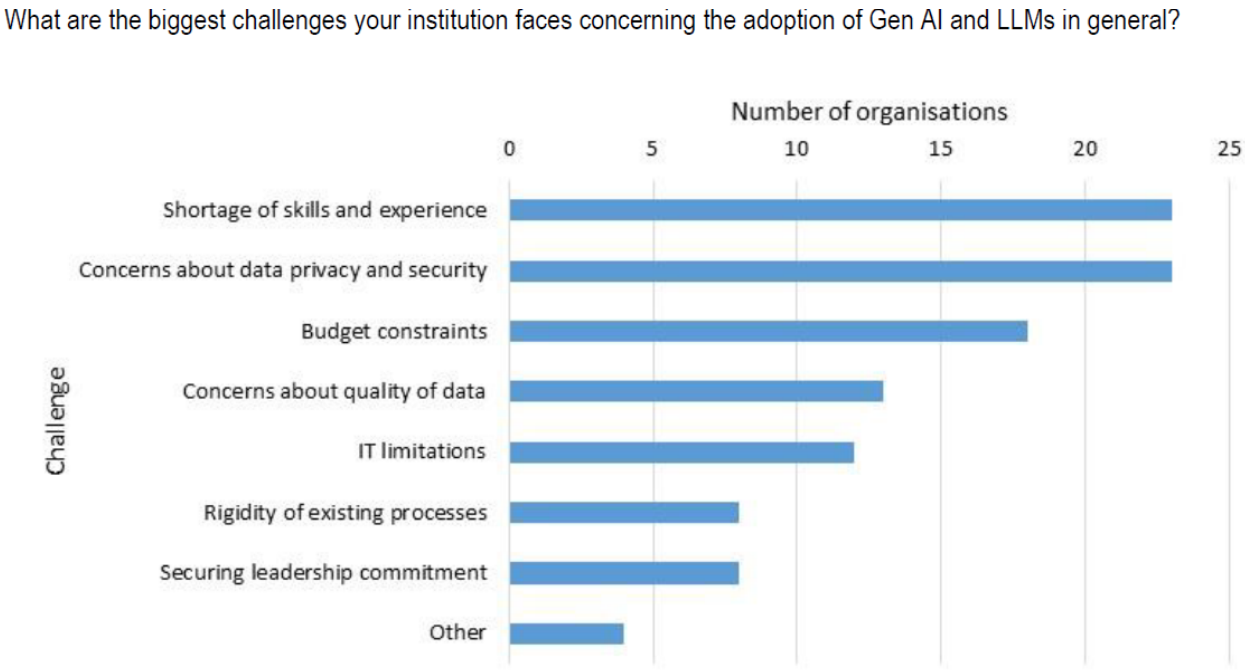
- authorities should publish plans for establishing or procuring AI systems in advance, with information on the purpose of those systems
- contracting authorities should require suppliers to demonstrate transparency (e.g. by making source code available to independent experts for periodic inspections)
- contracting authorities should have in place impact assessments and audit systems to reduce risks of misusing AI systems for private gain

Capacity building challenges

Alongside ensuring transparent and accountable deployment of AI in the anti-corruption field and developing the necessary regulatory and institutional frameworks, it is equally important to address capacity building challenges (OECD 2025).

A recent OECD survey of integrity actors found that a shortage of skills and experience ranked as the most pressing concern in implementing gen AI and LLMs, followed by concerns over data privacy and budget constraints (Figure 3).

Figure 3. Key perceived challenges for the adoption of gen AI and LLMs in the public sector.



Source: Ugale and Hall 2024: 26.

In relation to this, another pressing challenge is the mismatch between AI tools and institutional capacity, either because agencies lack the people/skills to use them effectively or because the tools themselves fail in ways that carry legal consequences. Brazil's Alice (Annex 1, Example 5) shows the first problem: chronic staff shortages left auditors overloaded with risk-alert emails, unable to triage everything, so potential red flags went unreviewed (Odilla 2023). The example from the UK Serious

Fraud Office illustrates the second: AI-assisted disclosure platforms (used to sift tens of millions of documents and identify legally privileged information at far lower cost compared to non-AI alternatives) later omitted material due to formatting/encoding issues, forcing reconfiguration and raising concerns that historical searches – and therefore disclosure in past cases – may have been incomplete (Ring 2024, 2025; Fisher 2025). In investigative settings, such failures can jeopardise prosecutions and undermine trust.

Some of these challenges can be addressed with careful planning from the onset and by learning from experiences of other countries. For instance, the findings of a recent OECD survey of using gen AI and LLMs in the public sector for anti-corruption and integrity suggest several ways of piloting and scaling gen AI initiatives (Ugale and Hall 2024: 29):

- start with incorporating gen AI into lower risk areas and processes (e.g. writing document summaries) as such an approach can help with capacity building in areas where mistakes are not that costly
- IT requirements need to be considered for both piloting and scaling AI initiatives, including computational and storage capacities, the availability of high-performance computing power, and data storage and data management capabilities

Further, the initiative with data and AI-driven audits in TdC in Portugal suggests the importance of strengthening data literacy and digital skills of TdC staff to ensure a successful implementation of these audits. This means that users of audit risk models need to undergo continuous training on how to use these tools, understand the outputs and interpret the results, as well as to be aware of any changes (e.g. introduction of new risk indicators) (Hlacs and Wells 2025: 23).

Annex 1: Examples of AI ACTs in corruption prevention

Example #1: Identifying strategies for limiting competition in public contracting (Katona and Fazekas 2024)

AI technology applied	Machine learning (logistic regression, random forest, XGBoost models) and classical NLP techniques.
Developed by/deployed by	Academics (Katona and Fazekas 2024).
Datasets used	All published government tenders in Hungary between 2011-2020 (approx. 119,000 contracts).
Main objectives	To show the relevance of textual information in bidding conditions, product descriptions and assessment criteria for predicting single-bidding in otherwise competitive markets.
Main benefits	Introducing contract-related textual information improved the model accuracy from 77% to 82% compared to their models containing only structured variables (e.g. no publication of the tender call).
Main challenges	A high rate of missing data, which is likely due to a lot of text being included in the full tender documents, rather than the official tender announcements. The authors used the latter as the first source, which had an irregular structure and contained difficult-to-process formats (i.e. scanned PDFs) (see Katona and Fazekas 2024). The study relies on only one proxy of corruption risk: single-bidding.

Example #2: Identifying politically connected firms (Mazrekaj et al. 2024)

AI technology applied	Machine learning (logistic regression, ridge regression, lasso, random forests and random forests with boosting).
Developed by/deployed by	Academics (Mazrekaj et al. 2024).
Datasets used	The dataset included a population of firms registered in the Czech Republic in 2018 (254,367 firms) filled with information on political donations, donating board members and those running for political office (from domestic sources and Orbis company database).
Main objectives	To show how machine learning techniques can be used to predict political connections.
Main benefits	Machine learning models accurately predict over 85% of politically connected firms based only on firm-level financial and industry indicators, suggesting the potential of ML for public institutions to identify firms whose political connections may represent conflicts of interest (Mazrekaj et al. 2024).
Main challenges	They use easily interpretable machine learning algorithms (i.e. random forests) rather than “black-box” models like neural networks. The trade-off is that the latter have even greater predictive accuracy, at the expense of explainability (Mazrekaj et al. 2024).

Example #3: Identifying public procurement cartels (Fazekas et al. 2023)

AI technology applied	Machine learning (logistic regression, random forests and gradient boosting machines).
Developed by/deployed by	Academics (Fazekas et al. 2023).
Datasets used	Data for 78 cartels in 7 countries between 2004-2021. Public procurement datasets from Bulgaria, France, Hungary, Latvia, Portugal, Spain and Sweden for the 2007-2020 period.
Main objectives	To test the predictive power of machine learning models to detect cartel behaviour in public contracting.
Main benefits	The models demonstrate that no single indicator – or even a small set of indicators – can reliably predict cartel behaviour. Instead, combining multiple indicators yields accuracy rates of 77–91% in predicting cartels across different countries. These models have clear policy relevance: they can support investigations and guide preventive interventions. For instance, given the finding that one-third of procurement markets in the seven selected countries are at high risk of cartelisation, the results could inform the design of preventive policies aimed at removing barriers to competition in these high-risk markets (Fazekas et al. 2023).
Main challenges	There are challenges with data quality of public procurement data (large amount of missing data and of certain fields, such as the losing bidder and bid price information). Learning models require adaptation and improvement with new learning materials (e.g. latest investigative results) (Fazekas et al. 2023: 27) to keep up with cartels changing their behaviour and tactics.

Example #4: Predicting cartel participants in public procurement contracting (Huber and Imhof 2023)

AI technology applied	Deep learning (CNNs) computer vision approach.
Developed by/deployed by	Academics (Huber and Imhof 2023).
Datasets used	Labelled procurement data from Japan and Switzerland with known episodes of cartel activity and competitive periods. These data allow the authors to build many pairwise "bid-image" samples for training and testing the CNN.
Main objectives	Test whether CNNs can reliably flag cartel participants from bidding patterns.
Main benefits	The CNNs on average correctly classify 19 out of 20 firms as cartel members or competitive bidders (Huber and Imhof 2023: 9). CNNs reach a high accuracy of 95% on average for both Japanese and Swiss bid rigging cases.
Main challenges	Variation across simulations, suggesting a need for a substantially larger pool of images for training and testing. Explainability due to the use of black-box models: CNNs are less interpretable, so human audit trails remain important in enforcement. Adaptation risk: publicly known indicators may be gamed; models require periodic retraining as strategies evolve.

Example #5: Alice (Análise de Licitações e Editais, Eng. Analysis of Biddings and Call for Bids) bot in Brazil

AI technology applied	Machine learning (mixed with text preprocessing, data mining and regular expression techniques) (see Odilla 2023: 386).
Developed by/deployed by	Public policy application: the office of the comptroller general and later improved by the federal court of accounts by including ML techniques in the existing regular expressions to conduct data mining on a daily basis (Odilla 2023; OPSI 2024).
Datasets used	The data comes from "tenders published on ComprasNet (public procurement portal), Official Gazette of the Federal Republic of Brazil, sanctions and ownership databases" (Odilla 2023: 363).
Main objectives	Scanning public procurement notices and contracts to spot risks and inconsistencies early, generating email risk alerts and data visualisation on the dashboard before the award of a public contract, so auditors can intervene (Odilla 2023).
Main benefits	The Alice helped suspend or cancel over R$9.7 billion (approx. US$1.75 billion) between 2019 and 2022 (OPSI 2024).
Main challenges	Auditors complained of being overloaded with emails and therefore unable to check every email alert due to staff shortages (Odilla 2023). Although auditable, it is not open to the public (Odilla 2023).

Example #6: Justina del Mar in Peru

AI technology applied	NLP.
Developed by/deployed by	Public policy application: WWF Peru, in collaboration with WWF US, artisanal fishers, shipowners, maritime authorities and subject-matter specialists.
Datasets used	The design of the tool included several phases. First, telephone surveys with artisanal fishers and shipowners to identify existing knowledge gaps in the sector, like reporting mechanisms. Second, the government was consulted about typical request topics from fishers and shipowners. Third, WWF Peru compared this information with their experiences in their work with fishing communities. Finally, this information was then systematised and additional data added from access to information requests and consultations.
Main objectives	To increase fishers' and shipowners' access to accurate, up-to-date regulatory information (WWF 2024). To reduce opportunities for petty corruption such as bribery, extortion or influence peddling in inspections and licensing (WWF 2024). To strengthen transparency and trust between artisanal fishing communities and authorities.
Main benefits	Accessibility: free, 24/7 availability via WhatsApp (WWF 2024). Community engagement: content co-created with target users increases trust and relevance. Ease of updates: regulatory or procedural changes can be quickly integrated into the chatbot's responses.
Main challenges	Dependence on technology access: requires mobile device access, internet connectivity, and WhatsApp literacy.

Example #7: Predicting public corruption in Spanish provinces (López-Iturriaga and Pastor Sanz 2018)

AI technology applied	Deep learning (neural networks: self-organising maps).
Developed by/deployed by	Academics (López-Iturriaga and Pastor Sanz 2018).
Datasets used	Cases of corruption reported by the media or went to the court in Spanish provinces between 2000 and 2012 (a database was gathered by the newspaper outlet El Mundo).
Main objectives	Demonstrating the utility of neural network approach to predict corruption in Spanish provinces based on economic factors.
Main benefits	The authors identified economic and political factors that increase corruption (e.g. taxation of real estate, the same political party remaining in power for a long time). The model could forecast corruption in some provinces up to three years in advance, offering space for preventive and corrective measures.
Main challenges	Data for corruption cases comes from cases reported by the media, introducing bias. Model opacity, due to the use of a black-box model.

Example #8: Predicting corruption related crimes in Italian municipalities (de Blasio et al. 2022)

AI technology applied	Machine learning (classification tree).
Developed by/deployed by	Academics (de Blasio et al. 2022).
Datasets used	Crime data for most Italian municipalities (8,049 out of 8,092) from the Ministry of Interior from which they take white-collar crime data for the period 2008-2014.
Main objectives	To predict white-collar crime rate (identified as crimes committed against public administration and public faith, as defined in the Italian penal code, which include corruption, bribery, embezzlement, abuse of authority and fraud).
Main benefits	Their models correctly classify over 70% of municipalities that experience an increase in corruption crimes.
Main challenges	The risk of bias in the crime data (e.g. possibility that corruption is more likely to be reported in some communities than in others).

Example #9: Early-warning models for predicting malfeasance in public procurement (Gallego et al. 2021)

AI technology applied	Machine learning (lasso and gradient boosting model).
Developed by/deployed by	Academics (Gallego et al. 2021).
Datasets used	Over 2 million public procurement transactions in Colombia between 2011-2015 from the government maintained database (the Sistema Electrónico para la Contratación Pública). They also used a database of vendors who received fines from the office of the comptroller general, those who were fined for breaching contracts, and data on contract amendments.
Main objectives	Testing the power of machine learning models to predict corruption investigations, breaches of contract or implementation inefficiencies related to public procurement contracts in Colombia.
Main benefits	Their models predict contracts that are likely to result in undesirable outcomes early on, which can be used by authorities to select which contracts should be audited. They identified a small number of characteristics of contracts that matter the most out of over 300 different features, including their size, duration and sector. Their models showed that variables related to contracts (their size and duration) were important predictors of malfeasance.
Main challenges	Their measures of malfeasance are prone to the "selective labelling problem" (Gallego et al. 2021: 6). The models require periodic adaptation to prevent corrupt actors from anticipating which traits are more likely to raise an alarm by the early-warning system (Gallego et al. 2021).

Example #10: Identifying collusion in public procurement contracting (García Rodríguez et al. 2022)

AI technology applied	Machine learning (linear models – stochastic gradient descent), ensemble methods (extra trees, random forest, adaBoost and gradient boosting), support vector machines, nearest neighbours (k-neighbours), neural network models (MLP), naïve bayes (Bernoulli naïve Bayes and Gaussian naive Bayes) and Gaussian process.
Developed by/deployed by	Academics (García Rodríguez et al. 2022).
Datasets used	Six public procurement datasets from Brazil (oil infrastructure projects), Italy (road construction), Japan (building construction and civil engineering), Switzerland (road construction, and road construction and civil engineering) and US (school milk market).
Main objectives	To show that machine learning techniques can detect collusion in public procurement contracting.
Main benefits	The study empirically demonstrates that ML tools can be implemented and be useful even when only few pieces of information are available from a large number of auctions, such as bid values and the winning bidder from each auction in their research. The analysis found that three models (extra trees, random forest and adaBoost) performed best overall. When complete information about auctions was available, these models correctly identified cartel behaviour in 81% to 95% of cases, with their accuracy holding up consistently across most datasets (with the exception of the US case).
Main challenges	Explainability issue, related to the use of some black-box models. Requirement for a large amount of reliable historical data, which may be challenging due to issues data availability.

Example #11: Predicting municipal level corruption (Gallego et al. 2022)

AI technology applied	Machine learning (random forests, gradient boosting machine, lasso) and deep learning (neural networks).
Developed by/deployed by	Academics (Gallego et al. 2022).
Datasets used	Disciplinary prosecutions conducted by the Colombian office of the inspector general in charge of monitoring the behaviour of public officials in the 2008-2011 and 2012-2015 mayoral periods.
Main objectives	To show whether machine learning can predict municipal level corruption, proxied by politicians' (i.e. mayors) misconduct in Colombia.
Main benefits	Based on 147 municipality level predictors grouped into ten categories (e.g. financial sector, conflict, human capital, local politics, public sector), their models show which categories of factors are best predictors of politicians' misconduct, including financial variables, local demographics, local politics and human capital variables, offering promise for identifying municipalities for prioritised audit (Gallego et al. 2022).
Main challenges	"Selective labelling problem", in other words the fact that the decision about who and when to prosecute is not random, and it can be politicised, and consequently is not free from bias (Gallego et al. 2022).

Example #12: Global Health Atlas (Transparency International Global Health 2025)

AI technology applied	Gen AI (LLMs).
Developed by/deployed by	Public policy application: Transparency International Global Health (2025).
Datasets used	A raw dataset had several hundred thousand news articles from around the globe about corruption manifestations in healthcare, with the data from Newscatcher. This was then filtered down to approximately 26,000.
Main objectives	To show the scale and complexity of integrity issues in health around the globe and counter corruption within health systems around the world (Transparency International Global Health 2025; n.d.).
Main benefits	Using AI to systematically identify and categorise public reporting of health-related corruption (Transparency International Global Health 2025) based on geographical location, integrity area (e.g. misappropriation), health area (e.g. medicines) and date range (Transparency International Global Health n.d.). The system identifies articles relevant to corruption, classifies the specific type of corruption (e.g., bribery) and extracts key metadata like location and date, creating the world's largest repository of evidence on health corruption.
Main challenges	The data only includes cases reported by the media, introducing risks of bias. The data is also affected by factors like freedom of the press, and saturation of internet news, among other challenges.

Example #13 Analysing lobbying records (Transparency International UK)

AI technology applied	Gen AI (LLMs), machine learning and NLP techniques.
Developed by/deployed by	Public policy application: Transparency International UK (n.d.).
Datasets used	Data sources include UK government transparency disclosures and the Scottish statutory lobbying register (Transparency International UK n.d.).[16] There were more than 140,000 lobbying records.
Main objectives	To automatically categorise lobbying meetings by analysing the meeting descriptions, hosts and lobbyists (Transparency International UK n.d.).
Main benefits	Using AI to categorise lobbying meetings into (currently) four categories: health, housing, defence and climate (Transparency International UK n.d.).
Main challenges	Occasional misclassification of a lobbying meeting.

[16] A paper based on the analysis of this data: Whiffen 2025.

Example #14 ALMA (Algorithm for Timber Legality in the Amazon, Peru)

AI technology applied	Machine learning (binary classification with Random Forest), integrated into a robust Drupal-based platform with Flask/Python services. It includes dynamic analytical dashboards, interactive risk maps, and auditable records.
Developed by/deployed by	Developed by the Environmental Governance Program of Proética (Peruvian chapter of Transparency International), with the support of the Environmental Investigation Agency (EIA). The platform features a reinforced architecture and advanced security measures for institutions and authorised users.
Datasets used	Data from the SIGO-sfc system of OSINFOR (field inspections), Forest Transport Guides (GTF), and access-to-information requests. These inputs feed the risk models and analytical dashboards.
Main objectives	- Analyse the information contained in GTFs to project the risk of illegality in timber products. - Provide interactive and objective tools for authorities, justice operators, buyers, the press, academia, and civil society to assess risks in the forest traceability chain. - Strengthen forest governance and support evidence-based decision-making to reduce illegal timber trade.
Main benefits	- Free and open access for multiple stakeholders. - Increased transparency in the forest sector. - Early identification of high-risk operations, enabling additional field verifications. - Technical credibility grounded in historical evidence and actual inspections. - Technological innovation applied to fighting corruption and environmental crimes. - Secure and scalable platform with role-based access. - Exportable and auditable records that reinforce institutional traceability.
Main challenges	- Continuous maintenance and updates required. - Dependence on the quality and completeness of input data. - Risk of users placing excessive trust in risk scores without conducting additional verifications through official sources.

Annex 2: Examples of AI ACTs in corruption detection

Example #1: Rosie in Brazil

AI technology applied	Machine learning: "a Python-programmed application that first applies hypothesis and test-driven development processes and then unsupervised learning algorithms to estimate a 'probability of corruption' based on standard deviations for each reimbursement receipt submitted by MPs" (Odilla 2023: 384).
Developed by/deployed by	Public policy application: Open Knowledge Brazil, as part of the civic-tech initiative Operação Serenata de Amor.
Datasets used	Public data on congressional spending and attendance, ownership registers and private data on Google, Foursquare and Yelp.
Main objectives	To detect and flag suspicious reimbursement claims by members of Brazil's congress. To publicly share suspicious cases to promote accountability. To empower citizens with easier access to structured transparency data via Jarbas, which is a platform that puts together data previously scattered across various government datasets, and which can be easily navigated (Cordova and Gonçalves 2019).
Main benefits	Automatically tweets suspicious cases, encouraging public scrutiny. Behavioural impact (reduced meal reimbursement spending by app. 10% after launch). Civic empowerment (enables citizens and journalists to verify and challenge expenses). Rosie identified 8,276 suspicious transactions and 735 different congresspeople involved in these suspicious transactions (Operação Serenata de Amor n.d.).
Main challenges	Fewer engagements on X than expected by creators (Odilla 2023: 379). A broader challenge of sustainability of accountability projects based on bottom-up initiatives (Odilla et al. 2022).

Example #2: Asset declaration system in Armenia

AI technology applied	Machine learning.
Developed by/deployed by	Public policy application: Armenia's corruption prevention commission.
Datasets used	Officials' asset declarations and other state databases.
Main objectives	To enable greater scrutiny of public officials and enhance accountability by making asset declaration information more accessible to state institutions and the general public (Harutyunyan 2023).
Main benefits	Streamlining data entry and collection. Automatically integrating data from other state agencies. Automated verification function comparing submitted data in new declarations with previously submitted declarations and other state databases. Identified discrepancies are marked as a red flag, triggering a comprehensive analysis of the official's assets. The public application programming interface (API) enables interested third parties to use their own software tools to sift through data on the public website. Initially, the algorithms for flagging corruption risks were based on a static set of corruption risk indicators, but plans are that after the initial trial period they will activate an ML component that would enable the system to learn from the data it processes, helping to identify novel corruptive and deceptive practices (Harutyunyan 2023: 4).
Main challenges	The algorithmic tool that is being developed to flag corruption risks is kept private, citing compliance with the EU's General Data Protection Regulation (GDPR) (Harutyunyan 2023: 5). The (current) reliance on static corruption risk indicators.

Example #3 DOZORRO (Transparency International Ukraine)

AI technology applied	Machine learning.
Developed by/deployed by	Public policy application: Transparency International Ukraine).
Datasets used	A database of public procurement tenders from Prozorro, an e-procurement system launched in 2016 (Transparency International Ukraine n.d.).
Main objectives	To detect and flag tenders with a high likelihood of corruption risk. To support CSOs in the DOZORRO network by prioritising which tenders to investigate.
Main benefits	Avoids static risk-indicator lists, reducing the chance officials can "game" the system. Identifies suspicious tenders before the contract execution. Proven to help prevent inefficient spending and uncover high-risk contracts (Transparency International Ukraine 2025).
Main challenges	Sustainability of bottom-up, CSO-led initiative.

Example #4: Zero Trust Programme in China

AI technology applied	Machine learning.
Developed by/deployed by	Public policy application (Chinese Academy of Sciences and the Chinese Communist Party's internal disciplinary bodies).
Datasets used	Over 150 central and local government databases, covering bank account records, property registries, corporate ownership records, land acquisitions and others (Chen 2019).
Main objectives	To detect irregular financial movements by civil servants. To identify conflicts of interest (e.g., relatives bidding for contracts). To calculate and score the probability of corrupt activity. To alert authorities for investigation if assigned scores exceed set thresholds.
Main benefits	High detection capacity: identified around 8,700 officials engaged in misconduct in pilot jurisdictions. Broad coverage of data: enabled cross-sector risk detection that traditional audits might miss.
Main challenges	Opacity: black-box decision-making, with limited ability to explain AI conclusions. Surveillance and privacy concerns: extensive surveillance of officials' financial and personal relationships. Political sensitivity: resistance from senior bureaucrats led to discontinuation in many areas.

Example #5: Identifying money laundering patterns (Gandhi et al. 2024)

AI technology applied	Machine learning (random forests, elastic net, lasso, gradient boosting, linear regression, logistic regression, k-nearest neighbours) and deep learning (MLP, CNN).
Developed by/deployed by	Academics (Gandhi et al. 2024).
Datasets used	Financial transaction data from the US Treasury's Financial Crimes Enforcement Network (FinCEN), with each record documenting a specific suspicious activity, covering the period 2014-2023 (some of the variables include the state that reported suspicious activity, year of reporting, industry sector, description of suspicious activity, etc.) (Gandhi et al. 2024: 11).
Main objectives	Testing the ability of ML models to predict variables from the financial transaction dataset.
Main benefits	ML models performed well in predicting financial transaction variables (state and year) and classifying credit and debit card activities. Random forest classifier had 99.9% accuracy in state prediction.
Main challenges	A major challenge in integrating ML models into AML (anti-money laundering) systems is the quality and availability of data, as financial transaction data can suffer from incompleteness, inaccuracies and bias (stemming from underreporting of suspicious activities, changes in reporting regulations and variations in data collection practices) (Gandhi et al. 2024: 24). Regulatory barriers and data privacy concerns may limit the access to financial transaction data (Gandhi et al. 2024: 24).

Example #6: Inspector AI (Evaluating suspicion transaction reports in Peru)

AI technology applied	NLP techniques using neural networks (text pattern recognition, reading free-text fields, applying filtering/scoring) (GIZ 2024; OAS 2023).
Developed by/deployed by	Public policy application: National University of Engineering in Lima and Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ).
Datasets used	Suspicious transaction reports.
Main objectives	To shorten the time for processing suspicious transaction reports, enabling the Peruvian financial intelligence unit to focus on analysis, considering that the body receives 23,000 suspicious cases annually, based on the most recent data (GIZ 2024).
Main benefits	The AI-assisted software automatically filters data, such as names, addresses and payment types from the suspicious transaction reports and reads free-text fields which previously could not be logged, running filters to flag high-risk cases. Concealed money laundering in relation to drug trafficking or corruption can be identified faster and charges brought faster (GIZ 2024). Since the software was introduced, the number of suspicious transactions reported to the public prosecutor more than doubled.
Main challenges	Explainability and auditability: the tool uses neural network NLP to structure and score suspicious transaction reports.

Example #7: Forest Foresight (predicting illegal deforestation)

AI technology applied	Machine learning.
Developed by/deployed by	Public policy application: WWF-Netherlands in collaboration with Boston Consulting Group, Deloitte, Amazon Web Systems and several academic institutions).
Datasets used	Historical geospatial data, satellite images (Radar, Sentinel 1) and socioeconomic variables (e.g. population density).
Main objectives	Once trained, it reads real-time satellite images and detects early deforestation predictors like expanding roads, and alerts local authorities (WWF 2023).
Main benefits	The tool, piloted in Gabon and Borneo, can predict illegal deforestation up to six months in advance with 80% accuracy (WWF 2022). In Gabon, it helped rangers detect an illegal gold mine, thereby protecting 74 acres of forest (WWF 2023). Key benefits relate to the ability of authorities to react proactively, reducing costs and avoiding irreversible losses (WWF Ecuador 2025).
Main challenges	Capacity building challenges: its effectiveness depends on the infrastructure and enforcement capacity on the ground.

Example #8: Identifying suspicious language patterns in documents

AI technology applied	Machine learning and NLP.
Developed by/deployed by	Public policy application: the European Anti-Fraud Office (OLAF).
Datasets used	Commission and member states' databases and various other sources (e.g. Panama Paper leaks).
Main objectives	Combining keywords that may signal corruption red flags in email exchanges (European Parliament 2021; Gerli 2024; Nicaise and Hausenkamph 2025).
Main benefits	Saves time, enables execution of complex tasks, automates processes.
Main challenges	Data challenges (missing values, lack of interoperability, etc.) (European Parliament 2021). Concerns about data bias (European Parliament 2021).

Example #9: SyRi[17] (system risk indication) in the Netherlands

AI technology applied	Machine learning.
Developed by/deployed by	Public policy application : deployed by the Dutch Ministry of Social Affairs and Employment.
Datasets used	The system relied on vast data sources, including work, fines, penalties, taxes, properties, housing, education, retirement, debts, benefits, allowances, subsidies, permits and exemptions, and others (Algorithm Watch 2020).
Main objectives	It was intended to detect social welfare fraud in the Netherlands by flagging individuals and neighbourhoods at high risk of welfare, tax or benefits fraud by analysing vast cross-referenced personal data, based on which it decided which citizens in the neighbourhood need to be investigated further (Algorithm Watch 2020).
Main benefits	In the first five years, five municipalities requested to analyse a neighbourhood, of which, only two of these projects were executed (Algorithm Watch 2020).
Main challenges	A lack of transparency for citizens about what happens with their data (Algorithm Watch 2020). According to the research of a Dutch newspaper De Volkskrant, none of the algorithmic investigations resulted in detection of new cases of fraud (Algorithm Watch 2020). A lack of algorithmic transparency: the Dutch government refused to provide the details on the models used. In 2020, a Dutch court ordered immediate halt of SyRi due to its violations of the Article 8 of the European Convention on Human Rights, which protects the right to respect for private and family life (Algorithm Watch 2020; Borgesius and van Bekkum 2021).

[17] Although this example refers to fraud, rather than corruption, it offers useful lessons for deploying AI ACTs as it raised significant privacy concerns with data handling, lack of transparency about the algorithm, as well as concerns about discrimination and bias as the tool was primarily used in low-income neighbourhoods (Algorithm Watch 2020).

Example #10: Illicit fishing activities globally (Paolo et al. 2024)[18]

AI technology applied	Deep learning (convolutional neural networks (CNNs)).
Developed by/deployed by	Academics (Paolo et al. 2024).
Datasets used	Satellite imagery between 2017-2021 covering over 15% of the ocean, in which over 75% of industrial activity is concentrated, and vessel GPS data (Paolo et al. 2024: 85).
Main objectives	To map industrial vessel activities and offshore energy infrastructure between 2017 and 2021 (Paolo et al. 2024).
Main benefits	They found that 72-76% of industrial fishing vessels do not appear in public monitoring systems, showing that substantial parts of activity are not publicly tracked (Paolo et al. 2024). Their mapping reveals potential hotspots of illegal fishing activity and can identify industrial fishing vessels encroaching on the artisanal fishing grounds (Paolo et al. 2024: 89). The dataset and technology are freely available.
Main challenges	Their study likely underestimates the concentration of vessels, particularly small vessels, due to data limitations. Risks of misclassification in high-traffic areas, like near cities in wealthy countries, e.g. misclassifying smaller craft as a fishing vessel.

[18] Although this example indicates criminal activity rather than corruption, the technology used has clear relevance for the anti-corruption field, as other cases in this Helpdesk Answer demonstrate the potential of satellite imagery to detect corruption risks.

Example #11: Detecting fake suppliers in public contracting (Wacker et al. 2018)

AI technology applied	Deep learning (convolutional neural networks (CNNs)).
Developed by/deployed by	Academics (Wacker et al. 2018).
Datasets used	A dataset of government suppliers with active contracts in 2016 and 2017 from the Brazilian public purchasing system and data about Brazilian companies from the Brazilian Federal Revenue Office about company addresses. They use Google Street View API to download images of company addresses.
Main objectives	To detect anomalies in government supplier images to identify fake suppliers by automatically distinguishing images of valid supplier locations from arbitrary buildings and landscapes.
Main benefits	High detection performance. By extracting deep features from a collection of Google Street View images using a pretrained convolutional neural network to classify supplier locations, the authors show the application of these features for identifying valid suppliers, regardless of the image perspective (i.e. the angle from which an image of a supplier building was captured) that was collected.
Main challenges	Narrow scope (it only detects whether a supplier building is non-existent based on visuals from Google Street View images). False positive and false negative cases, resulting from its inability to detect if there is a building at a supplier's address, which does not relate to the actual supplier; other kinds of fraud not related to buildings; that the supplier building looks like private residence, etc.

Example #12: DATACROSS I[19] (monitoring corporate ownership anomalies)

AI technology applied	Machine learning (logistic regression, naïve Bayes classifier, decision trees, bagged trees and random forests) and NLP.
Developed by/deployed by	Public policy application: DATACROSS I (2019-2021) was coordinated by Transcrime – Università Cattolica del Sacro Cuore, with the participation of the French anti-corruption agency (AFA), the Spanish police (CNP) and Investigative Reporting Project Italy (IRPI), with Bureau van Dijk as a data partner (Bosisio et al. 2021). The capabilities of the tool were expanded with DATACROSS II (2022-2024) covering 200+ countries and 300+ million firms, with more risk indicators and data sources (Datacross III n.d. a). A new expansion is envisioned in the current DATACROS III project (2024-2026) by integrating advanced AI-driven analytics to detect high-risk entities, while reducing false positives (Datacross III n.d. a). The data is further expanded to new financial and asset related data sources, including beneficial owner registers, company data, real estate data, public procurement data, virtual currencies and transactions, and satellite imagery (Datacross III n.d. b). In terms of AI technologies, the current project relies on ML, NLP based analysis of unstructured information, processing of satellite images and videos, and AI-driven anomaly indicators to detect illicit financial activities (Datacross III n.d. b).
Datasets used	The dataset of Bureau van Dijk - Orbis Europe and other sources (e.g. sanctions data and data on politically exposed persons for eight countries from LexisNexis WorldCompliance) to analyse the ownership structure of 56 million companies in 29 European countries.
Main objectives	Detect anomalies in firms' ownership structure that are indicative of a high risk of collusion, corruption and money laundering corporate ownership across EU countries (DATACROS I Project). Assess the distribution of opaque companies across EU jurisdictions and sectors.
Main benefits	Allows early detection of risk factors within legitimate companies[20] by complementing traditional approaches like searches of sanctions lists with ML algorithms that attribute risk scores to companies (Bosisio et al. 2021). All the considered ML methods showed satisfactory performance. With regards to logistic regression, the algorithms correctly

[19] The table assesses only the DATACROSS I project.

[20] "Legitimate companies" here refer to corporate entities operating in compliance with the law and not established for the purpose of committing or concealing criminal activity.

	predicted 83.3% of sanctions on companies, and 88% of sanctions on owners (Bosisio et al. 2021: 53). The tool complies with personal privacy and law enforcement procedures, as it was designed with the help of legal experts (Bosisio et al. 2021: 9). At the broader level, it enhances police investigations, especially with the cross-border cases, improves the effectiveness of cartel detection by competition authorities and enables investigative journalists, NGOs to monitor opaque interactions between business and politics (Bosisio et al. 2021: 13).
Main challenges	Relies on static risk indicators (developed based on company financial and ownership data). Requires retraining with new learning data to catch up with the emergence of novel illicit schemes.

Example #13: Enhancing public procurement audits in Portugal

AI technology applied	Deep learning.
Developed by/deployed by	Public policy application: the Tribunal de Contas (TdC), the supreme audit institution in Portugal.
Datasets used	Procurement data from various sources, such as the Competition Authority (AdC), the Portuguese transparency portal, and the Institute of Public Procurement, Real Estate and Construction (IMPIC) (Hlacs and Wells 2025: 19). 37 risk indicators related to public procurement were developed, with three approaches: i) rule based (e.g. contract signed before the award decision date); ii) inference based (e.g. ratio of the estimated value and contractual price); and iii) model based (collusion).
Main objectives	Enhancing early detection of public procurement irregularities through advanced data analysis and machine learning (ML) techniques[21] (Hlacs and Wells 2025: 16).
Main benefits	The expected long-term impact includes enhanced identification of risks and unusual transactions, improved early detection of potential irregularities and real-time monitoring, enabled by the application of ML techniques.
Main challenges	The vast majority of indicators are rule and inference based, making the audit risk model predominantly static. The labelled data for training was used from external sources, raising concerns about contextual differences, among other issues. Black-box model used, which is hard to interpret.

[21] As of yet, the audit risk model is not fully integrated within the TdC systems, and it is not a final version of the risk analysis model (Hlacs and Wells 2025: 20).

References

Aarvik, P. 2019. Artificial Intelligence: A Promising Anti-Corruption Tool in Development Settings? U4 Report 2019:1. U4 Anti-Corruption Resource Centre.

Adam, I. and Fazekas, M. 2018. Are Emerging Technologies Helping Win the Fight against Corruption in Developing Countries? Pathways for Prosperity Commission Background paper series No. 21, Oxford, United Kingdom.

AI21 Editorial Team. 2025. What is an Open-Source LLM? AI21Labs.

Algorithm Watch. 2020. How Dutch Activists Got An Invasive Fraud Detection Algorithm Banned.

Andersen, T. B. 2009. E-Government as an Anti-Corruption Strategy. Information Economics and Policy, Vol. 21(3): 201-210.

Association for Computing Machinery. 2017. Principles for Algorithmic Transparency and Accountability.

Berryhill, J. et al. 2019. Hello, World: Artificial Intelligence and Its Use in the Public Sector. OECD Working Papers on Public Governance, No. 36, OECD Publishing, Paris.

Boesch, G. 2023. What is Computer Vision? The Complete Guide. VISO AI.

Borgesius, F. Z. and van Bekkum, M. 2021. Digital Welfare Fraud Detection and the Dutch SyRI Judgment. IAPP.

Bosisio, A. et al. 2021. Developing a Tool to Assess Corruption Risk factors in firms' Ownership Structures – Final report of the DATACROS Project. Milano: Transcrime – Università Cattolica del Sacro Cuore.

Chen, S. 2019. Is China's Corruption-Busting AI System Being Turned off for Being Too Efficient? Tech in Asia.

Constantino, T. 2024. Women Make Up 29% Of The AI Workforce — Here's How To Fix It. Forbes.

Cordova, Y. and Gonçalves, E. V. 2019. Rosie the Robot: Social Accountability One Tweet at a Time. World Bank Blogs.

Datacross III. No date a. About.

Datacross III. No date b. Actions.

Davis, E. 2025. OpenAI Just Released a New ChatGPT That's 'Much Smarter Across the Board,' According to Its CEO. Here Are Some GPT-5 Prompts to Get You Started. Entrepreneur. 7 August.

De Blasio, G., D'Ignazio, A. and Letta, M. 2022. Gotham City. Predicting 'Corrupted' Municipalities with Machine Learning. Technological Forecasting and Social Change. Vol. 184.

Decarolis, F. and Giorgiantonio, C. 2022. Corruption red flags in public procurement: new evidence from Italian calls for tenders, *EPJ Data Science*, vol 11.

Dilmegani, C. 2025. Top 30+ NLP Use Cases in 2025 with Real-life Examples. AI Multiple Research.

Dorash, M. 2017. Machine learning vs. rule based systems in NLP.

European Commission (EC). 2019. A Definition of AI: Main Capabilities and Scientific

Disciplines. High-Level Expert Group on Artificial Intelligence.

European Commission (EC). 2024. AI Act Enters into Force.

European Investment Bank (EIB). 2025. Investigations Activity Report 2024.

Elbahnasawy, N. G. 2014. E-government, Internet Adoption, and Corruption: An Empirical Investigation. World Development, Vol. 57: 114-126.

Elbashir, M. 2024. EU AI Act Sets the Stage for Global AI Governance: Implications for US Companies and Policymakers. Atlantic Council.

European Public Prosecutor's Office (EPPO). 2025. Annual Report 2024.

Etzioni, A. and Etzioni, O. 2017. Incorporating Ethics into Artificial Intelligence. The Journal of Ethics, Vol. 21: 403-418.

European Parliament. 2021. Proceedings of the Workshop on Use of Big Data and AI in Fighting Corruption and Misuse of Public Funds - Good Practice, Ways Forward and How to Integrate New Technology into Contemporary Control Framework. Policy Department D for Budgetary Affairs Directorate General for Internal Policies of the Union.

European Parliament. 2023. EU AI Act: First Regulation on Artificial Intelligence.

Fazekas, M., Tóth, B. and Wachs, J. 2023. Public Procurement Cartels: A Large-Sample Testing of Screens Using Machine Learning. Working Paper series: GTI-WP/2023:02. Government Transparency Institute.

Fisher, J. 2025. Disclosure in the Digital Age: Independent Review of Disclosure and Fraud Offences. Home Office. UK Government.

Foti, J. 2025. Solving for Unknowns: AI for Government Accountability in Low-Data Environments. Medium.

Future of Life Institute. 2024. High-level Summary of the AI Act. EU Artificial Intelligence Act.

Gallego, J., Rivero, G. and Martínez, J. 2021. Preventing Rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement. International Journal of Forecasting, Vol. 37(1): 360-377.

Gallego, J., Prem, M. and Vargas, J. F. 2022. Predicting Politicians' Misconduct: Evidence from Colombia. Data & Policy, Vol. 4.

Gandhi, H. et al. 2024. Navigating the Complexity of Money Laundering: Anti–Money Laundering Advancements with AI/ML Insights. International Journal on Smart Sensing and Intelligent Systems, Vol. 17 (1).

García Rodríguez, M. J. et al. 2022. Collusion Detection in Public Procurement Auctions with Machine Learning Algorithms. Automation in Construction, Vol. 133.

Gerli, C. 2024. How Public Organisations Can Use AI in Anti-Corruption: What We Know So Far and Why We Need to Learn More About It. Brief. Hertie School, Centre for Digital Governance.

GIZ. 2024. Inspector AI Tackles Money Laundering.

Government of the Netherlands. 2024. The Government-Wide Vision on Generative Al of the Netherlands. Government of the Netherlands, The Hague.

Griffin, M. 2019. China's AI Anti-Corruption Program is Getting Shut Down for Being Too Good. 311 Institute.

Gurin, J. 2014. Open Governments, Open Data: a New Lever for Transparency, Citizen Engagement, and Economic Growth. SAIS Review of International Affairs, Vol. 34(1): 71-82.

Hariyani, D. et al. 2025. A Literature Review on Transformative Impacts of Blockchain Technology on Manufacturing Management and Industrial Engineering Practices. Green Technologies and Sustainability, Vol. 3(3).

Harutyunyan, H. 2023. Leveraging AI to Counter Corruption in Armenia In The Digitalization of Democracy: How Technology is Changing Government Accountability, Kerley, B. (ed). NED and FORUM.

Hausermann, H. et al. 2018. Land-Grabbing, Land-Use Transformation and Social Differentiation: Deconstructing "Small-Scale" in Ghana's Recent Gold Rush. World Development, Vol. 108: 103–14.

Herbert Smith Freehills Kramer. N.d. The Future of Disclosure: AI and Advanced Technology in the Criminal Disclosure Process.

Hlacs, A. and Wells, H. 2025. Using Digital Technology to Strengthen Oversight of Public Procurement in Portugal: the Use of Data Analytics and Machine Learning by the Tribunal de Contas. OECD Working Papers on Public Governance No. 83. OECD.

Hillsdon, M. 2024. From Forest-Listening to Advanced Remote Sensing, Can AI Turn the Tide on Deforestation? Reuters.

Holdsworth, J. and Scapicchio, M. 2024. What is Deep Learning? IBM.

Huber, M. and Imhof, D. 2023. Flagging Cartel Participants with Deep Learning Based on Convolutional Neural Networks. International Journal of Industrial Organization, Vol. 89.

IBM. No date. What is an AI Model?

IBM. 2021. What is Computer Vision?

IBM. 2023a. What are AI Hallucinations?

IBM. 2023b. Open Source Large Language Models: Benefits, Risks and Types.

InterpretML. No date. Explainable Boosting Machine.

Jambeiro Filho, J. 2019. Artificial Intelligence Initiatives in the Special Secretariat of Federal Revenue of Brazil.

Jenkins, M. 2021. Algorithms in Public Administration: How Do We Ensure They Serve the Common Good, Not Abuses of Power? Blog. Transparency International.

Jones, E. 2023. What is a Foundation Model? Ada Lovelace Institute.

Katona, E. and Fazekas, M. 2024. Hidden Barriers to Open Competition: Using Text Mining to Uncover Corrupt Restrictions to Competition in Public Procurement. Working Paper series: GTI-WP/2024:01. Government Transparency Institute.

Kossow, N. and Dykes, V. 2018. Embracing Digitalisation: How to use ICT to strengthen Anti-Corruption. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

Kossow, N. and Kukutschka, R. M. 2017. Civil Society and Online Connectivity: Controlling Corruption on the Net? Crime, Law, and Social Change, 1-18.

Kossow, N., Windwehr, S. and Jenkins, M. 2021. Algorithmic Transparency and Accountability. Transparency International Anti-Corruption Helpdesk Answer. Transparency International.

Köbis, N. 2023. Bribes for Bias: Can AI be Corrupted? Transparency International. Blog.

Köbis, N., Starke, C. and Rahwan, I. 2022a. The Promise and Perils of Using Artificial Intelligence to Fight Corruption. Nature Machine Intelligence, 4, 418–424.

Köbis, N., Starke, C. and Edward-Gill, J. 2022b. The Corruption Risks of Artificial Intelligence. Working paper. Transparency International.

Kucherenko, A. 2019. AI-Watchdog for Public Procurement. Medium.

Labbe, N. 2021. Detecting Illegal Gold Mining Sites in the Amazon Forest: Using Deep Learning to Classify Satellites Images. KTH.

Laforge, G. 2024. The Dangers of Imposing Global North Approaches to AI Governance on the Global South. Tech Policy Press.

Laurance, B. 2024. Roads of Destruction: We Found Vast Numbers of Illegal 'Ghost Roads' Used to Crack Open Pristine Rainforest. The Conversation.

Leech, G. et al. 2024. Ten Hard Problems in Artificial Intelligence We Must Get Right. arXiv: 2402.04464.

López Acera, A. 2023. Artificial Intelligence and the Fight against Corruption. Agència Valenciana Antifrau.

López-Iturriaga, F. J. and Sanz, I. P. 2018. Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. Social Indicators Research, Vol. 140(3): 975-998.

Martin, A. J. 2018. Serious Fraud Office Hires 'Artificial Intelligence Lawyer'. Sky News.

Maslej, N. et al. 2025. The AI Index 2025 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

Mate, E. 2025. Breaking It Down: Machine Learning, Deep Learning, Computer Vision, NLP, and Generative AI Explained. Medium. 24 January.

Mazrekaj, D., Titl, V. and Schiltz, F. 2024. Identifying Politically Connected Firms: A Machine Learning Approach. Oxford Bulletin of Economics and Statistics, Vol. 86(1): 137-155.

McGrail, S. 2023. What is White Box (Glass Box) vs. Black Box AI? TalentSelect.ai.

MinnaLearn. No date. Glass-Box and Black-Box AI – What's the Difference?

Mishra, H. 2025. What are Open Source and Open Weight Models? Analytics Vidhya.

Murel, J. 2024. What is Reinforcement Learning? IBM.

NetAppInstaclustr. 2025. Top 10 Open Source LLMs for 2025.

Nicaise, G. and Hausenkamph, D. S. 2025. Unlocking AI's Potential in Anti-Corruption: Hype Vs. Reality. Blog. U4 Anti-Corruption Resource Centre.

Nvidia. No date. What is Generative AI?

Organisation of American States (OAS). 2023. LV (Hybrid) Meeting of the Group of Experts for the Control of Money Laundering.

Odilla, F., De Figueiredo, V. and Dos Santos Veloso, C. 2022. Citizens and Their Bots that Sniff Corruption: Using Digital Media to Expose Politicians Who Misuse Public Money. SBAP.

Odilla, F. 2023. Bots against Corruption: Exploring the Benefits and Limitations of AI-Based Anti-Corruption Technology. Crime Law and Social Change, Vol. 80: 353-396.

Odilla, F. 2024. Unfairness in AI Anti-Corruption Tools: Main Drivers and Consequences. Minds and Machines, 34(3): 28.

OECD. 2019. AI Principles.

OECD. 2024. Recommendation of the Council on Artificial Intelligence.

OECD. 2025. Harnessing AI for Integrity: Opportunities, Challenges, and the Business Case Against Corruption. Business at OECD (BIAC) Anti-Corruption Committee Paper.

Open AI. No date. What is the ChatGPT Model Selector?

Open AI. 2025. GPT-5 is Here.

Open Stories. 2021. Through The Power of the People: Empowering Citizen Watchdogs.

Operação Serenata de Amor. No date. Numbers.

Observatory of Public Sector Innovation (OPSI). 2024. Robot Alice – Bid, Contract and Notice Analyser. OECD.

Paolo, F. S. et al. 2024. Satellite Mapping Reveals Extensive Industrial Activity At Sea. Nature, Vol. 625.

Pillay, T. 2024. The Gap Between Open and Closed AI Models Might Be Shrinking. Here's Why That Matters. Time.

Raghavan, M. et al. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 469–81. Barcelona Spain: ACM.

Ring, S. 2024. Lawyers Question UK Fraud Agency over Case Disclosure Problems. Financial Times.

Ring, S. 2025. AI Can Help Cut Time It Takes to Bring UK Criminal Cases, Says Review. Financial Times.

Rodríguez, M. J. G. et al. 2022. Collusion Detection in Public Procurement Auctions with Machine Learning Algorithms. Automation in Construction, 133.

Roy, T. 2023. The History and Evolution of Artificial Intelligence, AI's Present and Future. All Tech Magazine.

Royas, J.-P. F. 2017. Rolls-Royce in £671m Settlement of Bribery and Corruption Probe. Sky News.

Sanderson, C. et al. 2023. AI Ethics Principles in Practice: Perspectives of Designers and Developers. IEEE Transactions on Technology and Society, Vol. 4(2): 171-187.

Shim, D. C. and Eom, T. H. 2008. E-Government and Anti-Corruption: Empirical Analysis of International Data. International Journal of Public Administration, Vol. 31(3): 298-316.

Slagter, B. et al. 2024. Monitoring Road Development in Congo Basin Forests with Multi-Sensor Satellite Imagery and Deep Learning. Remote Sensing of Environment, Vol. 315.

Strawinska, M. No date. AI: A Game-Changer in Combating Corruption in Public Procurement. Butterfly Data.

Stryker, C. and Kavlakoglu, E 2024. What is Artificial Intelligence (AI)? IBM.

Stryker, C. and Holdsworth, J. 2024. What is NLP? IBM.

Syracuse University. 2025. Types of AI: Explore Key Categories and Uses.

TNRC. 2024. "Justina del Mar," a Virtual Ally to Prevent Corruption in the Artisanal Fishing Sector. TNRC Blog Post.

Transparency International. 2025. Addressing Corrupt Uses of Artificial Intelligence.

Transparency International Global Health. 2025. Press Release: New AI-Powered Global Health Atlas Unveiled to Expose Corruption in Health Systems.

Transparency International Global Health. No date. Health Atlas.

Transparency International UK. No date. Open Access UK.

Transparency International Ukraine. 2018. DOZORRO Artificial Intelligence to Find Violations in PROZORRO: How it Works.

Transparency International Ukraine. 2025a. April Results: Dozorro Helped Prevent Inefficient Spending of More than UAH 67 Million.

Transparency International Ukraine. 2025b. A July record: DOZORRO saves UAH 133 million for the budget.

Transparency International Ukraine. No date. Analytics for Procurement.

Ubaldi, B.-C. and Zapata, R. 2024. Governing with Artificial Intelligence: Are Governments Ready? OECD Artificial Intelligence Papers No. 20. OECD.

Ugale, G. and Hall, C. 2024. Generative AI for Anti-Corruption and Integrity in Government: Taking Stock of Promise, Perils and Practice. OECD Artificial Intelligence Papers, No. 12, OECD Publishing, Paris.

UK Government. 2025. Artificial Intelligence Playbook for the UK Government (HTML).

von Thun, M. 2023. Monopoly Power Is the Elephant in the Room in the AI Debate, Tech Policy Press.

Wacker, J., Ferreira, R. P. and Ladeira, M. 2018. Detecting Fake Suppliers using Deep Image Features. 7th Brazilian Conference on Intelligent Systems (BRACIS).

Wageningen. 2023. AI System Predicts Illegal Deforestation: Already Prevented the Clearing of 30 Hectares Near a Gold Mine'.

Wheeler, K. 2025. How Trump Scrapping AI Safety Regulations Impacts Global AI.

Whiffen, R. 2025. Power Politics: Emerging Insights into Climate Lobbying in Scotland. Transparency International UK.

WWF. 2022. One Step Ahead in Preventing Illegal Deforestation.

WWF. 2023. Could AI Help Stop Deforestation Before It Starts?

WWF Ecuador. 2025. Amazon At Risk: Forest Foresight, An Artificial Intelligence Tool that Predicts Deforestation Up to Six Months in Advance.

WWF. 2024. WWF Peru presented 'Justina del Mar', the Specialized Chatbot that Will Resolve Queries or Doubts About Inspection, Sanctions and Reporting Channels in the Artisanal Fishing Sector.

Zerilli, J. et al. 2019. Algorithmic Decision-Making and the Control Problem. Minds and Machines, Vol. 29: 555-578.

Zhao, H. et al. 2024. Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, Vol. 15(2)

Zinnbauer, D. 2025. Artificial Intelligence in Anti-Corruption – a Timely Update on AI Technology. U4 Brief 2025:1. U4 Anti-Corruption Resource Centre.

Disclaimer

All views in this text are the author(s)', and may differ from the U4 partner agencies' policies.

Creative commons

Corruption erodes sustainable and inclusive development. It is both a political and technical challenge. The U4 Anti-Corruption Resource Centre (U4) works to understand and counter corruption worldwide.

U4 is part of the Chr. Michelsen Institute (CMI), an independent development research institute in Norway.

www.u4.no

u4@cmi.no

U4 partner agencies

German Corporation for International Cooperation – GIZ

German Federal Ministry for Economic Cooperation and Development – BMZ

Global Affairs Canada

Ministry for Foreign Affairs of Finland

Ministry of Foreign Affairs of Denmark / Danish International Development Assistance – Danida

Norwegian Agency for Development Cooperation – Norad

Swedish International Development Cooperation Agency – Sida

Swiss Agency for Development and Cooperation – SDC

UK Aid – Foreign, Commonwealth & Development Office