# From Audit to Algorithm: Ethical Challenges of AI Inclusion in Public Tax Administration

Bhanu Pratap Singh
Chicago, IL, US

Gaurav Sehgal
St Louis, MO, US

## ABSTRACT
Globally, taxes are the unarguable lifeblood of a government body, instrumental in providing essential revenue to fund public welfare programs, building and maintaining critical infrastructure, and social welfare programs critical to societal stability and progress.[1] For decades, the public tax administration relied on human tax auditors to review returns, conduct interviews, and apply judgment within legal boundaries[2].

The fast-paced adoption of artificial intelligence (AI) in public tax administration resulted in unprecedented efficiency in revenue collection, fraud detection, and compliance monitoring [3][4]. A fast-changing department with some level of resistance to the change—from traditional human-led audits to algorithm-driven decision systems—it also raises profound ethical questions [5]. This article examines the transition through three lenses: **fairness** and **bias**, **transparency** and **accountability**, and **privacy** versus **surveillance**. Drawing on a few global case studies from the Netherlands, Canada, and India [6][7][8], this paper argues that while AI can bring efficiencies via reducing administrative costs and close taxation gaps and targets, unanswered ethical risks threaten public trust and democratic legitimacy [9].

Artificial intelligence (AI) promises transformative efficiency in tax administration, yet its deployment risks amplifying bias, eroding privacy, and undermining public trust if not guided by rigorous ethical safeguards. This paper proposes a policy framework rooted in human-centered values, fairness, transparency, robustness, and accountability—aligned with ISO/IEC 42001:2023 [11] and ISO/IEC 22989:2022 [12]to ensure AI serves taxpayers equitably while enhancing compliance and operational integrity. [11][12] Drawing on U.S. federal findings, international standards, and practical tools such as AI impact assessments, threat modeling (e.g., STRIDE), and the TRUST principles (Fairness, Accountability, Transparency, Privacy, Inclusivity), the framework outlines a lifecycle-based governance model tailored to tax contexts. [13][14] Key recommendations include mandatory bias audits, human-in-the-loop oversight for high-stakes decisions, public model registries, and regulatory sandboxes for low-risk testing. [15][16] An implementation roadmap with phased milestones and measurable KPIs demonstrates feasibility, illustrated through global benchmarks from Sweden, Australia, and Brazil. [17][18][19] By embedding these principles, tax authorities can harness AI's potential to reduce administrative burdens, minimize disparate impacts, and foster societal trust in digital governance.

## General Terms
AI Tax administration , AI Engineering, Machine learning, AI Governance, AI life Cycle, AI Threat Analysis, ISO standards for AI governance, Cloud Tools for AI governance, AI Risks Assessments, ISO 42001, ISO 31000, AI Bias Mitigation, AI Inspection, AI Transparency, AI Security and Privacy, AI Regulation and Legislations, AI Operation and Monitoring, AI System Retirement, Stride, Dread, OWASP security for AI, AI Risk Identification.

## Keywords
AI Tax administration , AI Engineering, Machine learning, AI Governance, AI life Cycle, AI Threat Analysis, ISO standards for AI governance, Cloud Tools for AI governance, AI Risks Assessments, ISO 42001, ISO 31000, AI Bias Mitigation, AI Inspection, AI Transparency, AI Security and Privacy, AI Regulation and Legislations, AI Operation and Monitoring, AI System Retirement, Stride, Dread, OWASP security for AI, AI Risk Identification.

## 1. INTRODUCTION
In the past, tax authorities mainly used human auditors. These are generally trained professionals who have reviewed the returns, conducted interviews, and applied their judgment within the law. However, these processes were not only vulnerable to human errors, but they also had problems with scalability and overall throughput.

Today, machine learning models have evolved to such an extent that they can even predict the risk of tax-related evasion, flag anomalies in real time, and, with minimal human intervention, initiate collection actions. For instance, the Internal Revenue Service (IRS) in the USA presently uses AI to evaluate more than 150 million annual tax returns. By doing so, it can prioritize cases of high risk with 85% accuracy (IRS, 2024 [20]). Moreover, similar systems at the same level of scale are working at Her Majesty's Revenue and Customs (HMRC [21]) in the UK and the Goods and Services Tax Network (GSTN [21]) in India where the use of machine learning algorithms is also at a very high level.

Moreover, taxpayers in the U.S. benefit from the country's AI leadership due to its diverse workforce [23]. To maintain this leadership, federal oversight must start with a review of existing regulatory frameworks [24] and focus on aspects such as bias mitigation, privacy protection, and accountability where there are gaps.

This paper presents a comprehensive policy framework named "Toward Ethical AI in Tax Administration" built around the main ideas of growth, sustainable development, human-centered values, fairness, transparency, explainability, robustness, safety, and accountability. Some of the U.S. policy discussion points that influenced this framework include the need for diverse AI teams, better access to non-sensitive government data, and risk-based regulation. Apart from that, the framework also adheres to global standards such as ISO/IEC 42001 [11] for AI management systems and ISO/IEC 22989 [12] for lifecycle management.

The TRUST principles play an important role in the structure. They imply actionable help: Fairness through mathematical metrics and continuous monitoring; Accountability through well-defined roles and legal safeguards; Transparency via model cards and explanation engines; Privacy secured by data

minimization and differential privacy; and Inclusivity to help overcome the digital divides. All this is made possible by regulatory sandboxes for controlled trials and congressional oversight, which allow gradual implementation from local trials to national scaling with measurement methods to ascertain the net benefits in terms of cost savings and trust. Globally, tax agencies' real-life instances make this framework a valuable tool for policy makers and managers to implement AI not only for improved productivity but also for maintaining fairness, ensuring taxpayer rights, and generating long-lasting societal trust in the digital transformation era.

## 2. FAIRNESS AND BIAS: PROTECTION OF ALGORITHMS TO INHERIT INEQUALITY

AI systems used in tax administration should behave only if the data they are trained on also is fair; however, most of the time, the data carry traces of the previously existing human biases. In general, the decisions on tax audits have been influenced by subjective judgments, limitations of resources, and enforcement priorities that targeted certain groups disproportionately over the last few decades. Consequently, when machine learning models learn from past audit records, they not only recognize noncompliance patterns but also become capable of inheriting and further prolonging discriminatory practices inherent in the system.

Various studies have demonstrated repeatedly that low-income filers, racial minorities, and small businesses are subjected to tax audits more frequently than other groups, though not necessarily that they commit tax fraud more often. The main reason why their tax returns are chosen for examination is that they are the easiest ones, while at the same time, it is improbable that there will be any costly legal disputes (Kleven et al., 2011 [25]; IRS SOI, 2022 [26]).

U.S. is a particularly good example of a problematic situation in this regard. Groundbreaking research conducted jointly by Stanford University, the Treasury Department, and other institutions showed that Black taxpayers are audited at rates 2.9 to 4.7 times higher than non-Black taxpayers (Ho et al., 2023 [59]). According to the research, most of the differential results from the algorithmic identifying of the returns, which take the Earned Income Tax Credit (EITC) route. EITC is a necessary refundable credit program that aims at providing help to low- and moderate-income working families. The differences in the audit rates remain even after the factors of income and error rates are controlled, which is a clear indication of proxy discrimination: the models seem to depend on indirect racial correlations such as location, family structure, or socio-economic variables hidden in the data. In 2024, a review of IRS practices held by the U.S. Government Accountability Office (GAO) led to the conclusion that, although the agency has already started working on understanding potential unintended biases in its selection processes, at present, oversight and data limitations hinder the realization of such risks in correspondence audits that are causing a heavy workload for vulnerable populations who find it difficult to react (U.S. GAO, 2024 [60]).

Canada demonstrated a similar pattern. The 2023 audit of the Canada Revenue Agency's AI risk-scoring system informed the auditor general of Canada that the postal code areas related to indigenous communities were flagged 40% more times than normally expected even after taking account of income and deduction patterns (Auditor General of Canada, 2023 [27]). In both scenarios, the respective algorithms did not explicitly use protected attributes like race or ethnicity but ended up reproducing discriminatory incidents through the indirect proxies.

The infringements compromised the fundamental notions of fairness: statistical parity (comparable error rates across groups) and individual fairness (treating similar cases similarly). If no conscious measure is taken - such as reweighting training data, implementing adversarial debiasing methods, or limiting proxy variable use - the systems might reinforce historical injustices which they have been programmed to handle, but they are doing so in a mathematically neutral manner.

Experiences from the real world illustrate that technical solutions alone will not suffice. Taxpayers should be given thorough reassurance that protective measures are in place and instructions on how to challenge unjust results. An increased public understanding of bias-related risks along with obligatory independent reviews, diverse dataset curation, and continuous supervision are all factors equally important in achieving this goal. Only from such steps can we escape the predicament of algorithms silently augmenting disparities when in fact they are supposed to help lessen them by implementing fair tax systems.

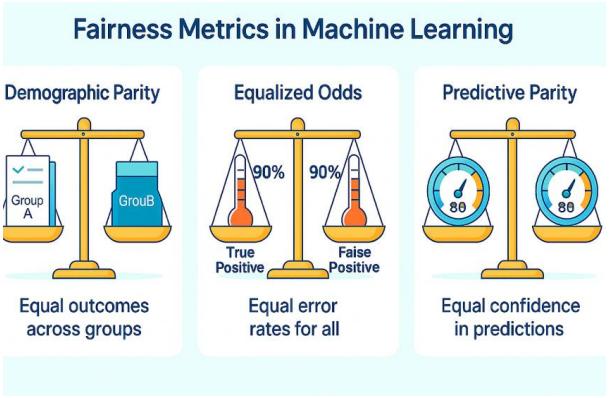


**Fig1: Fairness Metrics in ML**

## 3. TRANSPARENCY AND ACCOUNTABILITY: THE BLACK BOX PROBLEM

Taxpayers have a legal right to understand why they were audited or penalized. Yet many AI systems in use today are opaque even to their operators. This means even the closest team to maintain the system cannot provide certifications. The Dutch Belastingdienst's "Risk Prediction Model" (RPM), deployed between 2018 and 2022, was reportedly so complex that internal reviewers could not fully reconstruct its decision logic (Dutch State Audit Office, 2022 [29]). When challenged with legal proceedings in court, the agency admitted that they could not provide meaningful explanations for 12,000 contested cases.

This violation is not only unethical but also violates the EU's General Data Protection Regulation (GDPR [30]) Article 22, the article that grants citizens the right to be subject to fully automated decisions with legal effects—unless accompanied by human review and explanation.

Moreover, when such errors occur—who should be held accountable? The programmer? The data scientist? The tax commissioner? In traditional audits, responsibility is clear. In algorithmic systems, it diffuses across code, data, and policy.


**Fig2: Bias Vs Privacy balance**

## 3.1 Transparency and Explainability

The majority of AI models, especially deep neural networks, are mysterious "black boxes" whose inner workings are not clear. Citizens who are governed by algorithms have a right to be given reasons as part of their procedural rights (for instance, U.S. Fifth Amendment; EU GDPR Article [22]). Yet, proprietary algorithms in tax software often withhold model details, citing trade secrets.

The "explainable AI" (XAI) domain provides such a tool as SHAP (SHapley Additive explanations) values that can help to understand predictions (Lundberg & Lee, 2017 [31]). Nevertheless, simplified versions of explanations sometimes leave out the details, thus giving the user a false sense of understanding. In any case, transparency is one of the essential factors for public administration to keep the trust of the people: a 2022 Pew survey reveals that 68% of Americans say they would not trust a government AI decision if it was not transparent (Pew Research Center, 2022 [32]).

## 3.2 Accountability and Liability

Who can be called at fault when AI makes a mistake? Normal audits usually point the blame at human auditors; however, algorithm errors raise the issue of developers, data providers or agencies being responsible. In a tax scandal in the Netherlands 2021, AI inaccurately accused thousands of childcare subsidy fraud, thus making these families lose their money; the government had to resign as it failed to take the lead in the accountability process (Amnesty International, 2021 [33]).
Liability frameworks lag: contract law may shield vendors, while sovereign immunity protects agencies. This gap creates "accountability voids" (Floridi et al., 2018 [34]), where harmed taxpayers lack remedies.

# 4. PRIVACY VS SURVEILLANCE: THE HISTORICALK TAX STATE

Artificial intelligence is essentially driven by vast amounts of data, which modern tax systems are continually demanding. The current AI applications are not limited to basic tax returns only; they rather merge banking transactions, property records, utility bills, lifestyle indicators, and even social media activities to develop detailed behavioral profiles of the taxpayers. Out of these accounts, India's Project Insight, which was established in 2017, is a prime example of this pattern of operations by Data-mining cross-references more than 50 data sources to identify discrepancies and estimate noncompliance (Income Tax Department, 2024 [35]).

The main argument from the supporters is that this heavy data approach makes it possible to achieve accurate risk-based enforcement— detecting sophisticated evasion while refraining most honest taxpayers from the intrusive audits. However, the opponents warn that it is a very slippery path that may lead straight to predictive policing of daily financial behavior. For instance, if a taxpayer raises his charitable donations following a job loss, he may be the one to cause a "lifestyle inconsistency" alarm to go off although the activity is entirely legitimate and in full compliance.

The ethical tension around the issue is profoundly created by this:

- **Utility**: More data → better fraud detection
- **Cost**: Less privacy → chilled financial autonomy

The latest international regulations have those issues in mind. The 2024 technical note of the International Monetary Fund on AI in tax and customs administration that is dedicated to this subject, draws attention to the first legal and ethical obstacles arising from this direction, mainly in connection with the use of biometric data (such as facial recognition for identity verification) and large-scale personal data processing (IMF, 2024 [61]). Even though such innovations can facilitate the authentication procedures of the taxpayer and prevent identity fraud, they also evoke doubts about issues of proportion, as well as of consent and the risk of mass surveillance without any specific suspicion of a person.

According to the European Court of Human Rights rulings made on numerous occasions, the practice of mass and indiscriminate data retention conflicts with Article 8 rights of private life. At the same time, many tax AI systems deployed nowadays are population-wide profiling, which implies that they do not involve targeted investigations based on suspicion. In the absence of strong protective measures – data minimization, strict purpose limitation, regular independent audits, and obvious provisions for expiration of the retained data – these devices may be turning the tax administration function from a necessary civic one into an instrument of the state's pervasive oversight. The issue of enforcing the law effectively at the same time preserving the essential privacy rights is still among the most outstanding ethical problems of the algorithmic tax era.

# 5. CASE STUDIES: LESSION FROM THE FIELD

**Table 1. Describes the historic learning and the comparison from three case studies from the past.**

| Country | System | Outcome | Ethical Lesson |
|---|---|---|---|
| Netherlands | SyRI (2014–2020) | Halted by court for discriminatory risk scoring<br><br>[50] District Court of The Hague (2020)<br><br>[51] van Bekkum & Borgesius (2021) | Need for independent algorithmic impact assessments |

| | | | |
|---|---|---|---|
| Canada | CRA AI Risk Model (2021–) | Ongoing bias audits mandated<br><br>[52] Office of the Auditor General of Canada (2023) | Continuous monitoring > one-time fixes |
| India | GSTN Anomaly Detection | 300% rise in detections<br><br>[53] GSTN Annual Report (2023); [54] ResearchGate (2025) | High efficacy must not override due process |
| Australia | ATO AI Deployment (43 models in production as of May 2024, plus 8 approved public generative AI tools) | Enhanced compliance monitoring and taxpayer services; ongoing governance improvements (AI policy due Dec 2025)<br><br>[57] Australian National Audit Office (ANAO) Report No. 26 (2024–25) | Scaling AI with transparency (e.g., public registries, staff policies) balances innovation and accountability, but needs robust centralized oversight |
| United States | IRS Audit Selection Algorithms (ongoing, focus on EITC claims) | Black taxpayers audited 2.9–4.7 times more than non-Black (Stanford/Treasury research, confirmed 2023–2024)<br><br>IRS reduced EITC correspondence audits in FY2024 but GAO notes insufficient review of potential unintended bias<br><br>[55] Ho et al. (2023); [56] U.S. GAO (2024) | Race-blind systems can amplify disparities via proxy variables; require comprehensive data/model reviews and equity-focused reforms |

# 6. TOWARDS ETHICAL AI TAX ADMINISTRATION: A POLICY FRAMEWORK

The advantage of AI adaptation in tax administration leads to developing and practicing AI that is ethical, capable of reducing bias, promotes fairness, and protects privacy. These guiding principles are mandatory when it comes to aiming for a positive effect on society and building trust.

The inclusion steps for Principals of Artificial Intelligence in taxation needed to be based on pillars of growth, sustainable development and well-being, human-centered values and fairness, transparency and explainability, robustness, security and safety, and accountability [11].


**Fig3: Diagram to Ethical Pillars in AI driven administration**

## 6.1 Key Principles

### *6.1.1 Findings*

- The rise of artificial intelligence has offered potential to improve quality of life for taxpayers and taxation staff, provided it is developed and used in a manner that is ethical, reduces bias, promotes fairness, and protects privacy.
- A diverse workforce with an Artificial Intelligence skill set is required for Bias mitigation.
- The United States is uniquely positioned to leverage its diverse workforce to take a lead in artificial intelligence adaptation.
- The starting point for Federal oversight of artificial intelligence should be to review existing regulatory frameworks.
- Regulatory sandboxes, in general, refer to regulatory structures where a participant obtains limited or temporary access to a market in exchange for reduced regulatory uncertainty, and can be used to test a product designed to mitigate unintended bias or promote fairness in a small-scale environment and under the supervision of regulators.
- Federal programs should have necessary safeguards and oversight processes.

Artificial intelligence regulatory approaches should consider the level of risk associated with different artificial intelligence applications.

### *6.1.2 Bias Mitigation*

The agencies of the Federal Government should—

- Support technical and non-technical research and development to address potential bias, fairness, and privacy issues in artificial intelligence.

- Improve access to a broad range of non-sensitive government data assets to help train artificial intelligence systems.
- Implement title II of the Foundations for Evidence-Based Policymaking Act of 2018 (Public Law 115–435; 132 Stat. 5529).
- Develop policies to identify the data used to train artificial intelligence algorithms as well as data analyzed by artificial intelligence algorithms and systems in use by departments and agencies.
- Further develop and release to the public available benchmark data assets with the proper safeguards to protect privacy, mitigate bias, and promote inclusivity.

### *6.1.3 Regulation and Legislation Review*

- Review the range of existing Federal regulations and laws that potentially apply to artificial intelligence.
- Determine laws that apply to Artificial Intelligence.
- Determine if any gaps in appropriate legislation and regulation exist and how such gaps could be addressed.
- Advance Federal privacy reforms that build trust and prevent harm.
- Conduct regular oversight of artificial intelligence policies in the executive branch

The Congress should support funding for departments and agencies of the Federal Government interested in adopting programs, including regulatory sandboxes, for the purpose of testing artificial intelligence tools in limited markets.

With AI being a major part of the organizational strategy and operations, a proper and responsible AI governance based on trust is still very much needed. The main problem is how to keep AI systems ethical, tough, and in line with changes in regulations and industry standards.

ISO/IEC 42001 [11], the very first international management system standard for AI, has a full framework to help organizations figure out, set up and run an effective AI governance throughout the AI lifecycle. This article explores the role of ISO/IEC 42001 [11] in enabling trustworthy AI practices, briefly describes its core risk-management requirements, and provides examples of how threat modeling can be used as a practical method to fulfill these requirements.

## 6.2 AI GOVERNANCE

AI governance is about the organizational structures, policies, and controls that allow AI systems to be used in a socially responsible, ethical, and safe manner. Governance is a comprehensive concept that is applicable to the whole AI lifecycle and involves various activities such as:

- Defining the intended use and alignment with stakeholders
- Assessing the risks related to data, models, and deployment
- Implementing features for the system to be understandable, bias-free, and source able
- Creating practices for accountability, monitoring, and system removal
- These activities constitute the pillars of a formal framework which can be leveraged to setting up governance procedures, recognizing and handling risks, and putting into operation the processes for continuous improvement.

These initiatives, when combined, create the base of a well-organized governance structure that ensures justice, lessens the possibility of discrimination, increases the degree of responsibility, and facilitates the ongoing development of AI systems in general. With such a system in place, entities are empowered to recognize risks in a systematic manner, carry out the necessary protective measures, and keep up the openness at every stage of the AI technologies' evolution.

Such a framework constitutes a well-organized governance system grounded on the principles of justice, bias minimization, accountability elevation, and continuous improvement in AI systems. With this framework, organizations can methodically pinpoint risks, put in place the required safeguards, and practice openness during the entire AI technologies' lifecycle.

### *6.2.1 AI Life Cycle for Tax Administration*

Although ISO/IEC 42001 [11] lays down a firm and structured base for AI governance, ISO/IEC 22989:2022 [12], on the other hand, provides the features of AI systems and describes their development over time. In tax administration, it is pivotal to have stringent governance implemented in each stage of the AI lifecycle so as to be certain that AI-powered methods are the means through which fair tax enforcement, transparent decision-making, equal treatment of taxpayers, and system behavior accountability are kept.

Proper lifecycle management of the tax department means that the department can use a risk management strategy to foresee, understand, and reduce the risks arising from the taxes, especially those concerning bias, discriminatory outcomes, data quality, and taxpayer rights.

The lifecycle of AI as per ISO/IEC 22989:2022 [12] comprises various stages that are interrelated, and the stages are modified to show not only the changes in the responsibilities of a tax administration department but also the utility of such changes towards ensuring the fairness of the tax administration

#### *6.2.1.1 Inception*

The Inception phase is basically the point where a new AI idea in tax administration meets real life. At this stage, teams specify which business needs, e.g. by getting more accurate audit selections, improving taxpayer services, or recognizing complex fraud patterns, fulfilling compliance and service objectives most appropriately. Technical, legal, and resource constraints are also taken into consideration in feasibility studies so that the project can be in line with the organization’s capacity and requirements.

One of the main ethical issues leading to the focus on risk identification at the very beginning. When defining the problem, origin of bias (for instance, historically, due to enforcement disparities even certain income groups or regions, which were affected more than others) and risk of unfair treatment of taxpayers must be pinpointed concretely. Besides, stakeholder consultations with taxpayer representatives and civil society are carried out to consider equity and inclusion issues. By integrating bias risk checking from the very beginning - usually through first AI Impact Assessments (AIIAs) - agencies can avoid discrimination downstream and develop systems that are based on human-centered values rather than being a continuation of past inequities.

#### *6.2.1.2 Design and Development*

In the Design and Development stage, the AI system's architecture, data pipelines, and model specifications are the focus. It portrays an image of a defined framework of the

system, data flow, and sourcing of training and testing datasets while a few explicit guardrails are integrated to prevent algorithmic discrimination.

Fairness-by-design is a must: teams have to decide on and implement correct fairness measures (for instance demographic parity or equalized odds), set up strict documentation standards (like Model Cards) and, from the very beginning, build explainability features. Data quality checks are there to make sure the data is representative - by, for example, eliminating the situation where underrepresented groups of taxpayers are less represented due to historical records - and they also allow for techniques like reweighting or synthetic data generation to be used in the bias mitigation process.

Privacy and security provisions such as data minimization and differential privacy are put in place along with robust testing methods. Collaboration among different departments - data scientists, tax policy experts, legal advisors, and ethicists - helps the conversion of ethical principles into real technical decisions, thus providing a solid base for a trustworthy deployment.

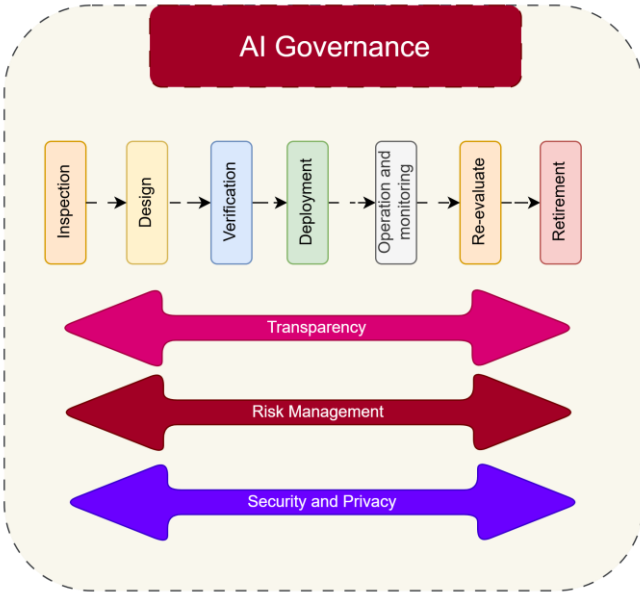


**Fig4: AI Governance lifecycle diagram illustrating the phases from inspection to retirement with three horizontal arrows below indicating transparency, risk management, and security and privacy throughout all stages.**

#### *6.2.1.3 Verification and Validation*

Testing AI models to check whether the models are up to the regulations, requirements, and moral standards of the operation, including fairness thresholds, transparency expectations, safety controls, and the giving up of disparate impacts on taxpayers. At the same time, validation ensures that the AI-partnered decisions are still interpretable and auditable.

#### *6.2.1.4 Deployment*

First step in a live tax administration environment of an AI system, with assuring tools for monitoring, human intervention, audit trails, and escalation mechanisms ready to help check and balance and the undergoing of the fair process by the taxpayer.

#### *6.2.1.5 Operation and Monitoring*

Using the AI system in the routine tax operations—such as audit selection, fraud detection, service automation, or collections—while always keeping a record of system functionality, checking model drift, and assessing the results in terms of fairness, correctness, and openness. Continued monitoring is a guarantee that no unintended discrimination or undue burden will arise for any taxpayer group.

#### *6.2.1.6 Re-evaluation*

Determining if the AI system is still in line with the objectives of the tax administration, the legal requirements, the ethical principles, and the expectations of the public. Re-evaluation also involves consideration of the model performance, taxpayer effect, and fairness metrics along with changes in laws, policies, and operational realities.

#### *6.2.1.7 Retirement*

The retirement of the AI system in a locally controlled and accountable way, making sure the tax data are correctly managed, the long-term retention and access are properly handled, and the audit logs kept for the support of openness and the accountability even after the system has been turned off.

Knowing and using this lifecycle is very important when it comes to foreseeing and lessening of the AI-specific risks within tax administration. The seven stages as laid down in ISO/IEC 22989:2022 [12] are however not carved in stone that each tax department can adjust or redefine the lifecycle to reflect its legal mandates, data environment, and governance responsibilities.

The stages of the lifecycle are the baseline for the next elements of an AI management system that include system scoping, threat modeling, fairness and bias risk assessment, transparency reporting, and continuous oversight throughout the broader AI governance program.

### *6.2.2 Risk Management in ISO/IEC 42001:2023 Framework for Tax Department*

Within a tax department, ISO/IEC 42001:2023 [11] provides a structured approach to managing the risks associated with AI systems used for taxpayer services, compliance activities, audits, and operational automation. After AI-related risks are identified and assessed, the department must implement appropriate operational controls to mitigate them. These controls—and the AI systems they support—should be continuously monitored, reviewed, documented, and improved to maintain reliability, fairness, and integrity across tax processes.

AI Impact Assessments (AIIAs) play an essential role when AI is used in high-risk areas such as fraud detection, automated case selection, predictive compliance, or eligibility determinations. AIIAs complement standard enterprise risk assessments by focusing on ethical, societal, and legal implications specific to tax administration. Similarly, Data Protection Impact Assessments (DPIAs), commonly required when processing large volumes of sensitive taxpayer information, help ensure compliance with privacy and data protection standards. Together, AIIAs and DPIAs create a holistic view of risks—covering fairness, transparency, privacy, accountability, and impacts on taxpayer rights.

Selecting the right AIIA tools or methodologies should be based on the specific AI use case—whether it is automated decision-making, data analytics, fraud detection, or case prioritization. Widely accepted frameworks include:

#### *6.2.2.1 ISO 3100*

An overarching risk management framework for business enterprises which supports revenue services in the recognition, assessment, and handling of operational, legal, and strategic risks in a systematic, consistent way. The framework is of

particular significance when the risks associated with AI are combined with other risks at the departmental level [36].

### *6.2.2.2 NIST AI Risk Management Framework (AI RMF)*

A specialized AI structure that focuses on explainability, robustness, fairness, and accountability - these being the main features of a public-sector tax system that needs to keep the trust of the public. The four core functions of the system (Map, Measure, Manage, and Govern) provide real help in the responsible implementation of AI in the taxation sector.

ISO/IEC 42001 [11] also underlines the necessity of detailed technical risk assessment through methods such as:

- **STRIDE**: This method helps to identify security threats that may lead to unauthorized access of taxpayer data or cause the interruption of essential tax operations.
- **DREAD**: This method is used to put first the threats that result in the unavailability of the system, the quality of data, or the delivery of services.
- **OWASP ML/AI Security**: This is a set of tools that helps the identification of vulnerabilities in machine learning models, and at the same time, it points to the risks of potential adversarial manipulation or the leaking of tax-related data [37].

The implementation of a reliable AI system in a tax department is the result of well-established governance practices, the presence of a clearly outlined risk management framework, and the performance of rigorous technical assessments. By bringing into line governance, operational controls, and technical safeguards, tax agencies can make sure that AI is a tool that promotes a fair, transparent, and legally compliant tax administration.

### *6.2.2.3 Threat Modelling for AI Risk identification*

With the help of threat modeling, the tax department can comprehend the technical risks that emerge from the AI lifecycle. These include things like attack surfaces, adversarial manipulations, and misuse scenarios, which go hand in hand with organizational risk assessments and impact analyses. By employing a threat modeling approach across the whole lifecycle, the department gets closer to conforming with ISO/IEC 42001:2023 [11] which is a standard for keeping AI as a technology for taxpayer services, compliance analytics, fraud detection, or automated decision making secure and trustworthy.

The below table presents a generic STRIDE threat model example for a generative AI resource in the tax department. It is divided by lifecycle stage and risk type. In addition, it indicates the ways in which standard cloud-native governance controls may be employed for remediation

### *6.2.2.4 Stride Threat Modelling*

**Table 2. AI Lifecycle and threat modelling table**

| Lifecycle Stage | Risk Types | Cloud Governance Controls |
|---|---|---|
| **Inception** (Spoofing) | Security | Identity and access management (IAM), multi-factor authentication (MFA), centralized authentication services, cloud-native threat detection tools |
| **Design & Development** (Tampering) | Integrity | API gateways, web application firewalls (WAF), policy-enforced guardrails, secure development pipelines, input auditing, immutable logging services |
| **Verification & Validation** (Repudiation) | Accountability | Comprehensive audit logging, AI model invocation logs, lineage tracking tools for prompts, outputs, and model versions |
| **Operation & Monitoring** (Information Disclosure) | Privacy, Security | Data anonymization tools, secure private networking, encryption in transit and at rest, strict data handling policies enforced through cloud configurations |
| **Deployment** (Denial of Service) | Availability | Rate limiting, automated scaling, load balancing, traffic filtering, DDoS protection services |
| **Re-evaluation** (Elevation of Privilege) | Ethics, Access Control | Role-based access control (RBAC), least privilege permissions, centralized policy enforcement, configuration monitoring, tamper-evident logs |

To reconcile innovation with ethics, tax authorities should adopt the following **TRUST** principles:

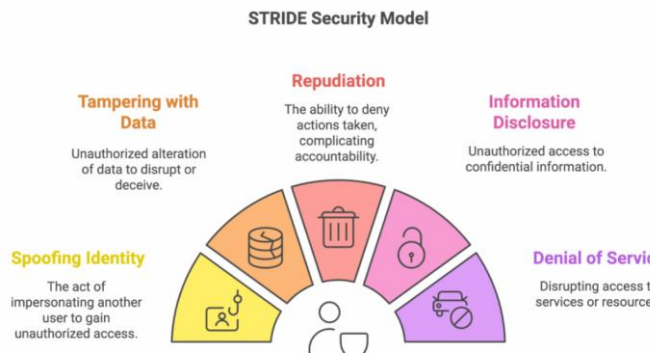


**Fig5: Diagram to showcase components of Stride Security Model**

## 6.3 AI GOVERNANCE PILLARS

### *6.3.1 Pillar 1: Fairness*

**Defining Fairness in Tax AI**

Mathematical fairness is not one-size-fits-all. Tax agencies must select from:

- **Demographic Parity**: Equal selection rates across groups
- **Equalized odds**: Equal true/false positive rates
- **Individual fairness**: Similar taxpayers treated similarly

Recommendation: Use **equalized odds** as default; adjust via stakeholder consultation.

#### *6.3.1.1 Implementation Steps*

- **Pre-deployment Bias Audit**
- Third-party review of training data and model outputs
- Mandatory reporting of disparity ratios (e.g., audit rate by income quintile, ethnicity)
- **Continuous Monitoring Dashboard**
- Real-time fairness KPIs published internally (and redacted versions publicly)
- **Redress Mechanism**
- "Fairness Challenge" portal: taxpayers can request re-evaluation if they suspect bias

*Example*: Sweden's Skatteverket now adjusts risk weights quarterly based on fairness drift detection (Skatteverket, 2025).

### *6.3.2 Pillar 2: Accountability*

**Table 3: Accountability Chain**

| Role | Responsibility |
|---|---|
| **AI Governance Board** | Strategic oversight, model approval |
| **Chief Algorithm Officer (CAO)** | Technical compliance, risk reporting |
| **Human Review Panel** | Final sign-off on high-impact actions (e.g., asset seizure) |
| **External Ethics Auditor** | Annual independent review |

### *6.3.3 Pillar 3: Transparency*

#### *6.3.3.1 Mandatory Disclosure Tool*

1. **Model Card** (Mitchell et al., 2019)
   - Purpose, data sources, performance, limitations
2. **Decision Explanation Engine**
   - Generates natural-language rationale (e.g., "Your return was flagged due to 300% increase in business expenses vs. industry norm")
3. **Public AI Registry**
   - Online database of all active tax AI systems

*Global Benchmark*: Australia's ATO publishes simplified model cards in 12 languages (ATO, 2025).

### *6.3.4 Pillar 5: Privacy*

**Table 4: Data Modernization Protocol**

| Data type | Retention | Access level |
|---|---|---|
| Tax returns | 7 years | restricted |
| Bank feeds | 90 days | AI |
| Social media | never | Prohibited |

#### *6.3.4.1 Privacy by Design*

- **Differential privacy** noise added to training datasets
- **Federated learning** where possible (e.g., regional
- models train locally, sharing only updates)

### *6.3.5 Pillar 6: Inclusivity*

#### *6.3.5.1 Bridging the Digital Divide*

- **Offline alternatives**: Paper forms, phone support, in-person centers
- **Digital literacy programs** funded via 0.1% of AI cost savings
- **Multilingual AI interfaces** (minimum: top 5 languages by taxpayer base)

*Case Study*: Brazil's Receita Federal reduced AI-triggered disputes by 28% after launching community tax clinics with AI explanation sessions (Receita Federal, 2025).

## 7. METHEDOLOGY

The rollout of the ethical AI framework that has been suggested is organized in phases, with the iterative roadmap balancing the technical aspects of the deployment with the ethical oversight that is continuous. The approach used here is influenced by risk-based regulation principles and features human-centered design so that taxpayers' point of view and societal values will be leading the way throughout the whole process.

A major element of this strategy is the establishment of participation methods which ensure that the different stakeholders - taxpayers, civil society groups, and advocacy organizations that represent low-income and minority communities, and independent ethicists - are involved in coming to the important decisions. In the Assessment phase (0-6 months), besides mapping the existing AI systems, structured consultations and focus groups are also conducted to gather qualitative insights on perceived risks, such as bias in audit selection or difficulties in redress. The voices of people help the baseline ethics audit by not only pointing at the technical side of things but also at the lived experiences of those who have been disparately affected.

The Governance phase (6-12 months) sees the formation of an independent Human Oversight Committee made up of external experts and taxpayer representatives. The group, among other things, scrutinizes the formation of roles, for example, that of the Chief Algorithm Officer and the public AI registry, thereby ensuring that accountability structures reflect democratic priorities and not just for the convenience of administration. The Piloting phase (12-24 months) makes use of regulatory sandboxes to implement the framework with limited context and with incessant feedback loops. The taxpayers who are affected by the pilot systems may express their views through surveys, community forums, and a special "Fairness Challenge" portal that enables the adjustment in real-time before the broader rollout.

The Scale & Evaluate phase (24+ months) serves to formalize the annual impact reports that are co-reviewed by the Oversight Committee and that incorporate long-term data of trust metrics, complaint volumes, and equity outcomes. The approach used here, which involves the embedding of human deliberation everywhere, rather than treating ethics as a compliance checkbox, is in line with the emerging worldwide best practices (OECD, 2025) and supports the TRUST principles, especially inclusiveness and accountability. In the end, it changes AI

governance from a technical top-down task into a collaborative one that maintains public trust in algorithmic tax administration.

**Table 5: Timeline Vs Objective Mapping Phase wise**

| Phase | Timeline | Milestones |
|---|---|---|
| **Assessment** | 0-6 months | Inventory all AI systems; conduct baseline ethics audit |
| **Governance** | 6-12 months | Establish CAO role, AI Board, public registry |
| **Piloting** | 12-24 months | Test FATPI in one region (e.g., VAT compliance) |
| **Scale & evaluate** | 24+ months | National rollout; annual impact reports |

# 8. EVALUATION MATRIX

To make sure that the suggested ethical AI framework results in real benefits without infringing on the rights of taxpayers, a strong evaluation matrix is required. This part of the document introduces the key performance indicators (KPIs) that reflect the progress of the TRUST principles: fairness, accountability, transparency, privacy, and inclusivity. Examples of such metrics are bias ratios in audit selection, satisfaction of the taxpayers with the decision explanations, the volume of privacy complaints, and overall trust results (revenue gains being offset against dispute costs) that give measurable proof of the success.

These signs allow continuous control, independent confirmation, and flexible management. Tax authorities by setting up challenging yet feasible targets like keeping bias ratios below 1.2:1 for different demographic groups and getting more than 80% explanation satisfaction, can show their accountability to the public and at the same time, they can justify their AI investments. Being regularly accountable against these KPIs enhances transparency and at the same time, it provides trust from the society in algorithmic tax administration.

**Table 6: Key Evaluation Metrics and Targets for the Proposed Ethical AI Framework in Tax Administration**

| Metric | Target | Measurement |
|---|---|---|
| Bias Ratio | < 1.2:1 across groups | Audit selection disparity |
| Explanation Satisfaction | > 80% | Post-decision taxpayer survey |
| Privacy Complaints | < 0.5% of AI cases | Ombudsman data |
| Cost Savings vs. Trust | Net positive | Revenue gain – dispute resolution cost |

# 9. ENABLE AI GOVENANCE & RISK MANAGEMENT WITH CLOUD TOOLS

The three major cloud providers in the US do provide tools that can be leveraged to build cloud governance frameworks. Below are the details

## 9.1 AWS tools for AI governance and risk management

The capabilities of AWS governance services support the control requirements specified in the Statement of Applicability (SoA) according to ISO/IEC 42001 [11]. These services and features empower companies to put into effect responsible AI principles on a large scale, and their usage is consistent with the content of ISO/IEC 42001 [11] which gives priority to the management of AI lifecycle in a structured and accountable way [38].

### *9.1.1 Amazon SageMaker Model Cards*

This is a service that offers uniform documentation for ML models covering aspects of purpose, performance, and limitations. From the point of view of governance, model cards serve as a tool to regulate the transparency, accountability, and auditability of the behavior and use of the model.

### *9.1.2 Amazon SageMaker Clarify*

*It* identifies bias in datasets and models and is a great help in the explainability of predictions. This is in line with governance rules that are directly related to the identification of fairness, non-discrimination, and explainability of AI models.

### *9.1.3 Amazon SageMaker Ground Truth*

Is a provider of data labelling workflows that are of high quality, with human involvement in the loop. It is a supporter of data governance as it ensures that the labelled datasets are accurate, consistent, and traceable.

### *9.1.4 Amazon Bedrock Guardrails*

Are they helpful in setting safety filters for generative AI like not allowing toxic content or that which is harmful. This then helps to align content and ethical governance policies.

### *9.1.5 AWS CloudTrail and AWS Config*

Are tools for audit logging and system changes monitoring on a continuous basis. These tools are a must for putting into practice the principles of accountability, traceability, and compliance reporting that are key elements of AI governance frameworks.

### *9.1.6 AWS Identity and Access Management (IAM)*

AWS Key Management Service (AWS KMS), and AWS Private Link: While IAM regulates access, AWS KMS handles encryption as well as key management, and Private Link allows for a confidential connection. In general, access control through access management, data security through encryption, and privacy through secure connections are the functionalities that these features are highly instrumental in achieving.

### *9.1.7 AWS Generative AI Lens*

It is one of the tools in the AWS Well-Architected Framework. It gives well-structured guidance for the assessment and enhancement of the architecture of generative AI systems. It assists organizations in the execution of responsible AI practices and the risk management process

## 9.2 GCP Tools for AI governance and Risk management

Google Cloud products dedicated to AI compliance and risk handling enable enterprises to bring to life responsible AI at large operational levels, which is quite in agreement with ISO/IEC 42001 [11] emphasis on the requirement of a well-structured, measurable, and traceable AI lifecycle. Like AWS, Google Cloud's offers are very much in line with the controls listed in the Statement of Applicability (SoA) of the standard and focus on transparency, fairness, security, and continuous improvement. I have listed the equivalent Google Cloud tools alongside the AWS ones, which you referred to, and gave a short explanation of their ethical AI deployment contribution [39].

### *9.2.1 Vertex AI Model Cards*

Formalized media for ML models, recording aspects like the goal, measured values, ethical implications, restrictions, and use case propositions. Monitoring governance mechanisms, they provide transparency and enable audit processes by granting the ability of the teams to trace the origin of the models and share the insight with all the stakeholders thus, making them more accountable.

### *9.2.2 Vertex AI Explainable AI (XAI) and Model Bias Metrics*

The instruments on the board used for determination of the most influencing factors (e.g., through SHAP or Integrated Gradients) to explicate the predictions and find the bias in the datasets or the outputs. This, as fairness controls, helps uphold identification of the early-stage biases in the datasets, facilitates the system explainability, and, if at any time during development or monitoring, the risk of discrimination arises, it is well-equipped to neutralize it.

### *9.2.3 Vertex AI Data Labeling and Human-in-the-Loop Workflows*

Fully serviced data labeling combined with an active learning system and consensus methods for data annotation which is of high quality and traceability. This is a great support for data governance as the program guarantees standardization, lessens the chances of error in the labeling, and introduces the human factor to oversee the ethical side of data throughout the AI lifecycle by integrating the human oversight feature.

### *9.2.4 Vertex AI Safety Filters and the Responsible AI Toolkit (for instance, SynthID Watermarking, Content Safety Classifiers)*

Generative AI's configurable guardrails, among which are automated filters for malicious content, toxicity detectors, and watermarking — assist in providing tracing for the outputs. The measures stipulated here are in line with the ethical aims since they prevent the potential misuse of the tech by stopping the biased or unsafe generation of outputs and thus indirectly also help producers in content moderation and the user community in providing safe content in the production environment.

### *9.2.5 Cloud Audit Logs and Cloud Monitoring*

The whole set of API calls, model installations, and changes in configurations, together with the alerts and dashboards for control in real-time, comprise a very complete monitoring system. It is very important to have such tools for traceability purposes, accountability, and the compilation of compliance reports. It is there where the automation of audits can be initiated to quickly identify drifts or irregularities in AI systems.

### *9.2.6 Cloud Identity and Access Management (IAM), Cloud Key Management Service (Cloud KMS), and VPC Service Controls*

IAM provides very detailed role-based access management, Cloud KMS ensures smooth handling of the whole chain in the life of the encryption key, and VPC Service Controls are there to provide connection that is private and secure. All these measures are aimed at implementing the privacy of data, access control, and protection rules, thus they act as a shield not only to AI training data but also to AI inferences preventing them from being accessed by unauthorized users.

### *9.2.7 Secure AI Framework (SAIF) Generative AI Lens*

A component of Google Cloud's Well-Architected Framework, it offers risk evaluation templates, best-practice guides, and assessment tools especially designed for generative AI. It helps the company in examining the designs from the point of view of their being robust, fair, and secure and thus it paves the way for responsible practice integration and lifecycle risk management in advance.

All these instruments are tightly interwoven with Vertex AI, which is Google Cloud's one-stop ML platform, and moreover, they are reinforced by Google's ISO/IEC 42001:2023 [11] certified AI management system. When it comes to putting this into practice, a good point to start would be with the SAIF Risk Assessment tool which you can use to gauge your present situation in relation to ISO controls. In case you are moving your operations from AWS, it will be easier and less problematic transition-wise with the help of good interoperability features Google Cloud has to offer (e.g., Big Query for data pipelines) while at the same time you will be able to maintain the same level of governance rigor.

## 9.3 Microsoft Azure tools for AI governance and Risk management

Microsoft Azure offerings for AI governance and risk management enable organizations to scale operationalization of responsible AI in a way that is compliant with ISO/IEC 42001 [11], which strongly supports not only managing but doing so in a controlled and accountable AI lifecycle. Just like AWS and Google Cloud, what Azure has to offer also helps the controls in the Statement of Applicability (SoA) under the standard, which are transparency, fairness, security, and continuous improvement, to be given the highest priority. To provide a reliable base for trustworthy AI, the AI Foundry Models and other components of Azure have been certified according to ISO/IEC 42001:2023[11]. In reference to your statement, I provide a comparison of Azure tools with the AWS ones and give the reasons how those tools help in ethical AI deployment.

### *9.3.1 Azure AI Model Cards*

A platform that offers uniform documents for describing ML models that clarify the purpose, performance metrics, ethical considerations, limitations, and intended use cases. These documents in governance processes help to bring transparency and auditability, as they allow the teams to monitor model lineage, inform the stakeholders and make sure that the whole AI lifecycle is accountable.

### *9.3.2 Azure AI Fairness and Interpretability Toolkit (part of Azure Machine Learning)*

A set of tools for identifying bias in datasets and models, and features of explainability such as SHAP-based feature importance to get the most understandable prediction. This, as a matter of fact, is a direct enabler of fairness control as it

reveals early on the discriminatory patterns, hence it is supportive of non-discrimination activities and gives the possibility for the explanation of the development and monitoring phases.

### *9.3.3 Azure Machine Learning Data Labeling*

Data-labeling and annotation workflows are performed in a human-in-the-loop fashion with the help of active learning and collaboration features. By making sure the labeled datasets are not only accurate and consistent, but also traceable, data governance is reinforced and, at the same time, human oversight is integrated to maintain the set ethical standards throughout the AI lifecycle.

### *9.3.4 Azure AI Content Safety (with Prompt Shields and Grounded Ness Detection)*

Methods to ensure the safety of user prompts in generative AI include prompt injections, jailbreaks, hallucinations, and detection of harmful content like toxicity and hate speech. From the ethical point of view, the existence of these safety features can be justified because they help the system to obey content restrictions, it becomes almost impossible that a biased or unsafe output will be generated, and also production environments get the tools they need for regulating activities.

### *9.3.5 Azure Monitor, Application Insights, and Azure Policy*

They provide a wide-angle view which includes logging of API calls, model deployments, and configuration changes, together with alerting, dashboards, and policy enforcement for the oversight that is done in real-time. In fact, they are very important, especially when it comes to traceability, compliance reporting as well as accountability. Moreover, they allow the execution of automated audits, whose task is to locate drifts, anomalies, or possible violations of the existing AI systems policies.

### *9.3.6 Azure Active Directory (Entra ID), Azure Key Vault, and Azure Private Link*

Entra ID is used to implement fine-grained role-based access control, Key Vault is for managing the full lifecycle of the encryption keys, while Private Link offers a safe and private way of connecting. The ones that are responsible for enforcing data privacy, access governance, and protection standards are great because they prevent unauthorized users from accessing sensitive AI training data and inferences while at the same time, they come to shield these from eavesdropping, tampering, or interception.

Microsoft Azure Well-Architected Framework is the Responsible AI Dashboard together with the Microsoft Responsible AI Standard Tools, which provides a set of risk assessment templates, fairness metrics, causal analysis, and evaluation checklists for AI systems. The principal aim of the platform is to guide enterprises through the complex tasks of identifying, measuring, and handling risks, e.g., bias or security threats, by applying the most suitable methods for leading to the use of responsible practices and conducting lifecycle governance proactively.

Microsoft AI Studio and Azure Machine Learning, the comprehensive AI platform of Microsoft, unify all these instruments contributed by Azure along with the ISO/IEC 42001:2023 [11] certification for the primary services such as Azure AI Foundry. To take the first steps, the Responsible AI Dashboard can be used for the automated fairness assessments of the models. In the case of a move from AWS or Google Cloud, Azure hybrid compatibility (e.g., through Azure Arc for multi-cloud) would make it not only easier but also more efficient while still being able to maintain the governance level.

## 10. FUTURE SCOPE

The next 10 years will see the growth of AI in tax administration at an incredible pace. This growth will be largely due to multimodal models, generative AI, quantum-resistant cryptography, and global real-time data ecosystems. Although the ethical framework and governance instruments proposed in this paper serve as a solid base for current systems, new technologies and changes in society will require constant adjustments. Ultimately, it is not about increasing efficiency on a larger scale; rather, it is about creating systems that are adaptive, can anticipate, and are deeply trustworthy thus maintaining democratic legitimacy in a time when algorithmic governance is ingredients.

## 10.1 Generative AI and Conversational Tax Administration

Generative AI, including large language models and multimodal foundation models, will be with us in the back-office for analytics only until 2030–2035. From then on, their interaction with taxpayers will be direct:

- In principle, AI tax assistants can hold natural language conversations and understand numerous languages and dialects.
- Imagine auditors who are virtual, in real time and aware of the context, capable of not only explaining the decisions but also negotiating the payment plans and even simulating “what-if” scenarios for tax planning.
- Automated creation of personalized nudges for compliance as well as of educational material.

New ethical issues: among them the risk of hallucination in tax advice, deepfake fraud (for example voice cloning of a taxpayer), and the potential of generative AI being weaponized by the bad actors for creating synthetically tax evasion. Governance of the future should, therefore, not only cover prompt injection defenses, output grounding verification, AI-generated watermarking, and the stipulation of human intervention in the case of any generative output with legal effect but also include extension of ISO/IEC 42001 [11] and 22989 for these aspects.

## 10.2 Real-Time Global Data Sharing and Borderless Compliance

By the end of the 2020s, measures like the OECD's Crypto-Asset Reporting Framework (CARF), Pillar Two real-time reporting, and centralized beneficial ownership registries will have woven a global network of financial data. Tax authorities will no longer rely on annual filings but will engage in continuous transaction monitoring:

- AI systems will be able to process live bank feeds, blockchain ledgers, decentralized finance (DeFi) activity, and IoT-generated economic signals in real-time
- Cross-border collaborative models wherein national tax AIs exchange risk signals without infringing on sovereignty or privacy

To accomplish this, there must be federated learning between different jurisdictions, homomorphic encryption, and zero-knowledge proofs to ensure that privacy is maintained while still allowing for the combined intelligence to be used against multinational profit shifting and crypto tax evasion.

## 10.3 Quantum Computing and Post-Quantum Cryptography

Quantum computers are expected to be able to break the encryption that is used today (RSA [48], ECC [49]) around the middle of 2030s. This will be a threat to the confidentiality of the tax data that will have been kept for decades. Tax agencies, therefore, have to move to quantum-resistant algorithms (NIST post-quantum cryptography standards) and get their AI systems ready which in the near future may also be running on quantum or hybrid quantum-classical hardware—thus, speeding up the detection of fraud patterns but at the same time, there being a possibility of completely new attack vectors.

## 10.4 Personalized "Tax Equity as a Service"

Tomorrow's systems may change the way we look at punitive audit selections by turning them into proactive equity ones:

- An AI that is always updating effective tax rates across the whole society to keep the tax burden from falling too heavily on the less privileged.
- Behavioral nudges at the individual level help close the compliance gaps among the economically disadvantaged groups without penalizing them.
- Linkage with universal basic income or social welfare schemes for instant benefit and liability adjustments.

This conception imparts tax management as a means of real-time economic justice rather than a revenue function. However, this would work only if issues of bias, surveillance, and lack of autonomy had been sorted out already.

## 10.5 Citizen-Centric and Participatory Governance Models

New paradigms being considered are:

- Public model registries changing to open-source or community-audited tax AI modules
- "AI citizens' assemblies" in which representative taxpayer panels collaboratively design the risk-scoring criteria
- Decentralized identity along with verifiable credentials giving taxpayers the power to decide the exact data points to be shared with tax authorities

## 10.6 Institutional Evolution Required

The ethical framework put forward in this paper should ultimately be an evolving one to stay sound morally and valid in time:

- Set up permanent Ethical AI Observatories devoted to monitoring by tax authorities, universities, and society at large
- Require five-year sunset clauses for all AI systems in the tax domain that are of high risk unless they are re-certified by up-to-date impact assessments
- Draft an international treaty on ethical tax AI under OECD or UN umbrella—comparable to the Geneva Conventions for algorithmic taxation
- Invent a continuing learning program to be certified tax officials, (like medical boards) merging technical, legal, and philosophical aspects of the training

# 11. CONCLUSION

The shift from audits to algorithms in tax administration is not merely technical, it is a governance transformation. AI can close revenue gaps and modernize public finance, but only if designed with ethical foresight. Left unchecked, it risks creating a tax system that is efficient for the state but unjust for the citizen.

Managing AI risk effectively means aligning technical, organizational, and ethical considerations throughout the AI system lifecycle. ISO/IEC 42001 [11] provides structure and accountability. Threat modeling techniques such as STRIDE, MITRE ATLAS, and OWASP [37] for LLM surface deep technical risks. AWS services and features such as SageMaker Model Cards, SageMaker Clarify, and Amazon Bedrock Guardrails help embed governance into layers of AI development.

By combining technical tools, structured assessments, and standards-driven controls, you can build AI systems that are trustworthy, resilient, and aligned with societal expectations.

For additional guidance on achieving, maintaining, and automating compliance in the cloud, contact AWS Security Assurance Services (AWS SAS [41]) or their account team [40]. AWS SAS is a PCI QSAC and HITRUST Assessor Firm that can help by tying together applicable audit standards to AWS service specific features and functionality. They help you build frameworks such as ISO 42001, PCI DSS [42], HITRUST CSF [43], NIST-CSF [44] and Privacy Framework, SOC 2 [45], HIPAA [46], ISO 27001 [47] and 27701, and more.

# 12. REFERENCES

[1] OECD (2024). Tax Administration 2024: Comparative Information on OECD and other Advanced and Emerging Economies.

[2] IMF (2023). Digitalization of Tax Administration.

[3] World Bank (2024). GovTech Maturity Index – Tax Administration Module.

[4] Gupta, S. et al. (2023). AI in Tax Systems: Opportunities and Risks. IMF Working Paper.

[5] Zorn, K. et al. (2024). Ethical AI in Public Sector. Government Information Quarterly.

[6] van Bekkum, M. (2022). The Dutch SyRI Case. European Law Journal.

[7] Auditor General of Canada (2023). Use of AI in Tax Risk Scoring.

[8] Income Tax Department India (2024). Project Insight Annual Report.

[9] European Parliament (2024). AI Act and Public Administration.

[10] Executive Office of the President (2023). Blueprint for an AI Bill of Rights.

[11] ISO/IEC 42001:2023. Artificial intelligence — Management system.

[12] ISO/IEC 22989:2022. Artificial intelligence — Lifecycle processes.

[13] NIST (2023). AI Risk Management Framework 1.0.

[14] Microsoft (2023). STRIDE Threat Model.

[15] UK Government (2024). Regulatory Sandboxes for AI in Public Services.

[16] OECD (2023). Recommendation on AI Governance.

[17] Skatteverket (2025). Fairness Monitoring Quarterly Report.

[18] Australian Taxation Office (2025). Public AI Registry.

[19] Receita Federal do Brasil (2025). AI Dispute Reduction Report.

[20] IRS (2024). Annual Data Book FY 2023–2024.

[21] HMRC (2024). AI and Automation Strategy 2024–2027.

[22] GSTN India (2024). Analytics and AI Roadmap.

[23] National Science Foundation (2024). Diversity in AI Workforce Report.

[24] Financial Conduct Authority (2024). Regulatory Sandbox Cohort 10 Results.

[25] Kleven, H. J. et al. (2011). Unwilling or Unable to Cheat? Econometrica, 79(3).

[26] IRS Statistics of Income (2022). Audit Rate Disparities by Income.

[27] Office of the Auditor General of Canada (2023). Report 5.

[28] Barocas, S. et al. (2023). Fairness and Machine Learning. MIT Press.

[29] Dutch State Audit Office (2022). Algorithmic Risk Prediction in Tax Administration.

[30] Regulation (EU) 2016/679 (GDPR), Article 22.

[31] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS.

[32] Pew Research Center (2022). Americans' Views on Government Use of AI.

[33] Amnesty International (2021). Netherlands Childcare Benefits Scandal.

[34] Floridi, L. et al. (2018). AI4People — An Ethical Framework. Minds and Machines.

[35] Income Tax Department India (2024). Project Insight Phase III Report.

[36] ISO 31000:2018. Risk management — Guidelines.

[37] OWASP (2024). Top 10 for LLM & AI Security.

[38] AWS (2025). ISO/IEC 42001 Certification Scope.

[39] Google Cloud (2025). Secure AI Framework (SAIF) v2.

[40] AWS Security Assurance Services. (2025). *AWS Security Assurance Services Overview*. Amazon Web Services.

[41] AWS. (2025). *AWS PCI DSS Compliance Responsibilities*.

[42] Payment Card Industry Security Standards Council. (2023). *PCI DSS v4.0*.

[43] HITRUST Alliance. (2025). *HITRUST CSF v11*.

[44] NIST. (2024). *Cybersecurity Framework v2.0 & Privacy Framework v1.0*.

[45] AICPA. (2024). *SOC 2 Trust Services Criteria*.

[46] U.S. Department of Health and Human Services. (2024). *HIPAA Security Rule*.

[47] ISO/IEC 27001:2022 & ISO/IEC 27701:2019. *Information Security and Privacy Information Management*.

[48] Rivest, R. L., Shamir, A., & Adleman, L. (1978). A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM, 21(2), 120–126.

[49] Koblitz, N. (1987). Elliptic Curve Cryptosystems. Mathematics of Computation, 48(177), 203–209.

[50] District Court of The Hague. (2020). Judgment in case C/09/550982/HA ZA 18-388 (SyRI legislation in breach of European Convention on Human Rights). ECLI:NL:RBDHA:2020:865.

[51] van Bekkum, M., & Zuiderveen Borgesius, F. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security, 23(3), 1–18.

[52] Office of the Auditor General of Canada. (2023). Report 5: Use of AI in Tax Risk Scoring.

[53] Goods and Services Tax Network (GSTN). (2023). Annual Report 2022–23. Ministry of Finance, Government of India.

[54] Leveraging AI Tools for Enhanced GST Compliance and Fraud Detection in the Indian Taxation System. (2025). ResearchGate Publication.

[55] Ho, D. E., et al. (2023). Measuring and Mitigating Racial Disparities in Tax Audits. Stanford Institute for Economic Policy Research Working Paper.

[56] U.S. Government Accountability Office (GAO). (2024). Tax Enforcement: IRS Audit Selection Processes for Returns Claiming Refundable Credits Could Better Address Equity (GAO-24-106126).

[57] Australian National Audit Office (ANAO). (2025). Auditor-General Report No. 26 2024–25: Governance of Artificial Intelligence at the Australian Taxation Office.

[59] Ho, D. E., et al. (2023). Measuring and Mitigating Racial Disparities in Tax Audits. Stanford Institute for Economic Policy Research Working Paper.

[60] U.S. Government Accountability Office (GAO). (2024). Tax Enforcement: IRS Audit Selection Processes for Returns Claiming Refundable Credits Could Better Address Equity (GAO-24-106126).

61] International Monetary Fund (IMF). (2024). Understanding Artificial Intelligence in Tax and Customs Administration. IMF Technical Notes and Manuals. Available at: https://www.imf.org/en/Publications/TNM/Issues/2024/xx/xx/Understanding-Artificial-Intelligence-in-Tax-and-Customs-Administration.