

**Artificial Intelligence in Tax Administration: A Conceptual Examination of Automation,
Compliance, Revenue Generation and Challenges Facing Modern Tax Systems**

Arowolo, Isiaka O (Ph.D)

Department of management and Accounting,
faculty of Management and Social Science,
Lead City University.

Email: arowologbenga@yahoo.com

Abstract

Tax administration occupies a distinctive position in the broader literature on artificial intelligence (AI) in public administration: it is simultaneously among the sectors where AI adoption is most advanced, with major revenue authorities operating dozens of production AI models, and among the sectors where AI's failures have produced some of the most consequential legal reversals of automated government decision-making to date. This paper offers a conceptual examination of AI's role in tax administration that treats both facts as central rather than treating the second as an unfortunate footnote to the first. Integrating the economic deterrence model of tax compliance, the slippery-slope trust-and-power framework, and agency theory with recent evidence on AI-enabled automation, predictive risk profiling, and taxpayer-facing natural-language systems, the paper argues that AI strengthens tax administration's capacity to detect non-compliance and streamline routine processing considerably more reliably than it strengthens the trust relationship on which voluntary compliance, and therefore sustainable revenue generation, actually depends. Drawing on documented failures of automated tax and welfare decision-making, including Australia's Robodebt scheme and the Netherlands' System Risk Indication program, the paper shows that AI-enabled compliance systems can simultaneously become more accurate in the aggregate and less legitimate in the particular, with direct consequences for the voluntary-compliance foundation the slippery-slope framework identifies as tax administration's most efficient long-run outcome. The paper develops an integrated model connecting automation, compliance, and revenue generation through a governance layer addressing explainability, oversight, and redress, and considers the implications of that model for tax administrations, taxpayers, and future research.

Keywords: artificial intelligence; tax administration; tax compliance; automation; revenue generation; algorithmic bias; taxpayer rights; explainability.

Introduction

Tax administration has used computational tools to support enforcement since at least 1977, when McCarty's TAXMAN system offered an early demonstration that legal reasoning about tax liability could be partially formalized and automated. What has changed since then is not the basic premise but its scale and sophistication: contemporary revenue authorities increasingly deploy machine learning, natural language processing, and generative AI across taxpayer registration, return processing, risk profiling, audit selection, and taxpayer-facing advisory services. The Australian Taxation Office alone reported forty-three AI models in production as of May 2024, alongside eight generative AI tools approved for public use, processing tens of millions of documents to identify entities warranting further scrutiny. This scale of deployment places tax administration among the most AI-intensive domains of public administration anywhere in the world, and the evidence available on its effects is correspondingly more developed than in most comparable public-sector contexts.

That evidence is considerably more mixed than either enthusiastic or purely cautionary accounts of AI-enabled tax administration typically acknowledge. On one hand, the efficiency and detection gains are real and increasingly well documented: Zheng, Xu, Liu, Shi, Wang, and Dong (2024), surveying data-mining techniques applied to tax risk detection, find machine-learning approaches substantially improving the identification of returns warranting audit relative to the rule-based screening methods they replace. On the other hand, the same period has produced two of the most consequential legal reversals of automated administrative decision-making on record. Australia's Robodebt scheme, an automated system that matched welfare and tax records to identify overpayments, relied on an averaging methodology that ignored genuine income fluctuation and improperly shifted the burden of proof onto individuals; a Royal Commission ultimately found the scheme unlawful. The Netherlands' System Risk Indication program was found by the District Court of The Hague to violate taxpayers' privacy rights specifically because the risk profiles it generated could not be adequately explained. Guglyuvatyy (2025), examining both cases alongside the broader regulatory landscape, argues that these failures were not implementation accidents but structural consequences of deploying automated decision-making in a domain whose legal and institutional frameworks were not designed to accommodate it.

This paper develops a conceptual account of why these two facts — substantial detection and efficiency gains alongside consequential legitimacy failures — belong together rather than being treated as separate storylines. Its central argument is that **AI strengthens tax**

administration's detection and processing capacity considerably more reliably than it strengthens the trust relationship between taxpayer and state that classical and behavioral tax-compliance theory both identify as the foundation of efficient, low-cost revenue collection. Where that trust relationship deteriorates because AI-driven decisions cannot be adequately explained or contested, the resulting shift toward enforced rather than voluntary compliance, in the sense the slippery-slope framework distinguishes, may increase administrative cost and taxpayer resistance even as detection technology itself continues to improve. The paper reviews the theoretical foundations relevant to this argument; examines AI's contribution to automation, compliance, and revenue generation in turn; documents the specific challenges the Robodebt and SyRI cases illustrate; and develops an integrated model connecting these elements through the governance conditions this paper argues determine which outcome ultimately prevails.

Conceptualizing AI's Role in Tax Administration

The OECD (2025), reviewing digitalization and digital transformation initiatives across member tax administrations, frames AI adoption in this domain as continuous with a longer digitalization trajectory running from electronic filing through data-matching to predictive analytics, rather than as a discontinuous new phase. This framing is useful for the present paper because it discourages treating AI-enabled tax administration as an undifferentiated novelty and instead invites decomposition into the specific functions AI actually performs. Belahouaoui and Attak (2024), conducting a systematic textometric review of the digital-taxation literature published between 2016 and 2023, identify three recurring functional clusters across this literature: automation of routine administrative processing, predictive analytics applied to compliance risk, and taxpayer-facing digital services — a typology this paper adopts as its organizing structure for the sections that follow.

Sutton, Holt, and Arnold (2016), tracing three decades of AI research in accounting more broadly, find that predictions of AI's transformative effect on rule-governed professional domains have circulated since the 1980s expert-systems era, with actual adoption proceeding more gradually and cumulatively than early commentary anticipated. Tax administration is a partial exception to this general pattern: McCarty's (1977) TAXMAN experiment shows tax law was among the earliest domains AI researchers attempted to formalize, precisely because tax liability determination combines the kind of explicit, rule-governed structure early AI systems could handle with high-stakes consequences that made the formalization exercise worth attempting. This paper treats this early and sustained attention as evidence that tax

administration's relationship with AI is genuinely distinctive among public-administration domains, not simply a recent instance of a broader public-sector digitalization trend.

This early and sustained attention also means tax administration offers an unusually long empirical record against which to evaluate claims about AI's trajectory in rule-governed public-sector domains. McCarty's (1977) own reflections on TAXMAN, published in the Harvard Law Review rather than a computer-science venue, anticipated several of the governance concerns this paper develops at length below — in particular, the difficulty of fully formalizing legal reasoning that depends on context, discretion, and purposive interpretation rather than mechanical rule application. That a legal scholar working with 1970s-era AI technology identified this difficulty nearly five decades before Robodebt and SyRI made it concrete and consequential suggests the underlying tension this paper examines is not a novel feature of contemporary machine learning specifically, but a structural feature of applying any sufficiently capable automated reasoning system to a domain that combines rule-governed structure with irreducibly discretionary elements — a tension this paper argues has simply become harder to ignore as AI capability has grown from TAXMAN's narrow experimental scope to the forty-plus production models a major revenue authority like the Australian Taxation Office now operates.

Theoretical Foundations

The Economic Deterrence Model of Tax Compliance

Allingham and Sandmo's (1972) foundational model treats the decision to comply with tax obligations as a rational calculation under uncertainty: taxpayers weigh the benefit of underreporting income against the probability of detection and the penalty if caught, choosing the level of compliance that maximizes expected utility. This deterrence framework supplies the most direct theoretical rationale for AI-enabled detection technology specifically: if compliance is fundamentally a function of perceived detection probability, then AI systems that genuinely increase detection probability, as the evidence reviewed later in this paper documents, should straightforwardly increase compliance by the model's own logic. Subsequent research has found the Allingham-Sandmo model's predictive accuracy limited on its own terms — actual observed compliance levels are considerably higher than the model predicts given realistic detection probabilities and penalties in most jurisdictions — a gap the behavioral tax-compliance literature has worked to explain through factors the purely economic model omits.

The Slippery Slope Framework: Trust and Power

Kirchler, Hoelzl, and Wahl (2008) address this predictive gap directly, proposing that tax compliance results from the interaction of two distinct dimensions: the power tax authorities exercise, encompassing audit probability and penalty severity in the tradition of the deterrence model, and the trust taxpayers place in tax authorities, encompassing perceived fairness, competence, and legitimacy. Their central theoretical claim, the slippery-slope framework's namesake insight, is that compliance can be sustained through either dimension but that the two produce qualitatively different kinds of compliance: enforced compliance, sustained primarily through power and deterrence, tends to be minimal, resentful, and administratively costly to maintain; voluntary compliance, sustained primarily through trust, tends to be more complete, more stable, and considerably cheaper to administer, since it requires less continuous enforcement pressure to sustain.

This distinction supplies this paper's central theoretical lens for evaluating AI's net effect on tax administration. AI-enabled detection technology operates almost entirely on the power dimension of the slippery-slope framework: it increases the effective probability of detection, which the deterrence model and the power dimension of the slippery-slope framework both predict should increase compliance. But the evidence reviewed later in this paper, particularly the Robodebt and SyRI cases, suggests AI-enabled enforcement can simultaneously damage the trust dimension, if the systems generating enforcement actions cannot adequately explain or justify their determinations to the taxpayers affected. On Kirchler et al.'s framework, this is a theoretically coherent but practically costly outcome: an administration can increase power-based deterrence while eroding trust-based voluntary compliance, arriving at a compliance equilibrium that is more enforced and less voluntary than before AI adoption, even if aggregate detection accuracy has genuinely improved.

Agency Theory and the Taxpayer–State Relationship

Jensen and Meckling's (1976) agency theory, developed to explain the relationship between corporate owners and managers, extends usefully to the taxpayer-state relationship when tax administration is understood as a monitoring mechanism the state deploys to constrain the divergence between what taxpayers might otherwise report and what tax law actually requires. This framing positions AI-enabled detection as a monitoring-cost-reducing technology in the classical agency sense: cheaper, more comprehensive monitoring should, all else equal, reduce the residual loss from non-compliance that agency theory treats as the

ultimate cost borne by the principal, here the state acting on behalf of the broader taxpaying public. But agency theory's own logic also anticipates the trust problem the slippery-slope framework identifies from a different angle: monitoring mechanisms that taxpayers perceive as illegitimate or unaccountable can themselves generate a distinct category of cost, as taxpayers who no longer trust the monitoring system's fairness may increase avoidance behavior specifically in response to the monitoring itself, a dynamic classical agency theory, developed for corporate rather than civic relationships, did not need to theorize in this specific form.

Explainability by Design in High-Stakes Administrative Decisions

Rudin (2019) draws a distinction directly relevant to the governance failures this paper documents: between building supplementary tools that explain a complex AI model's behavior after the fact, and choosing, at the design stage, a model that is inherently interpretable, even at some cost to raw predictive accuracy. Rudin argues the first approach, post-hoc explanation, is systematically unreliable for high-stakes decisions specifically because post-hoc explanations can be unfaithful to what the underlying model actually computed, creating an appearance of accountability without its substance. This distinction reframes the SyRI and Robodebt failures in a way this paper treats as theoretically significant: both systems were criticized specifically for their inability to adequately explain individual determinations, which on Rudin's analysis reflects a design choice — opaque, complex modeling favored over interpretable alternatives — rather than an unavoidable cost of automating tax and welfare determinations. Applied to tax administration specifically, where the legal requirement to provide reasons for adverse determinations is well established in most jurisdictions' taxpayer-rights frameworks, Rudin's framework suggests explainability should function as a design constraint from the outset of system development, not as a feature retrofitted after a system's opacity has already produced legal or public-trust harm.

AI and Task Automation

The most straightforward and best-evidenced contribution AI makes to tax administration concerns the automation of routine, high-volume processing tasks. Cooper, Holderness, Sorensen, and Wood (2019), examining robotic process automation across Big Four accounting firms, find RPA deployed extensively in tax-preparation and compliance-support workflows, with firms reporting substantial reductions in the time required for

reconciliation and routine filing tasks — evidence directly relevant to tax administration, since much of the automation logic transfers between tax preparation on the taxpayer side and tax processing on the administration side. Choi and Xie (2026), studying generative AI use inside real accounting workflows more broadly, document measurable reallocation of professional time away from routine data entry toward higher-value review and communication activity, a pattern this paper expects to generalize to tax-administration contexts given the structural similarity of the underlying tasks.

Bruckner and Coil (2024), examining AI adoption specifically within modern tax agencies, document a similar pattern at the institutional level: agencies report AI's clearest and most reliably realized benefits in processing efficiency — faster return processing, more consistent anomaly flagging, reduced manual review burden — with more ambitious applications, such as fully automated discretionary determinations, remaining considerably rarer and more contested. This pattern is consistent with the task-based automation logic developed in the broader AI-and-labor literature: automation succeeds most reliably on well-specified, rule-governed tasks, which describes routine tax processing considerably better than it describes the judgement-intensive determinations — penalty mitigation, reasonable-cause assessment, complex entity classification — that have historically required substantial discretionary judgement from tax officials.

Workforce Implications of Tax Administration Automation

Acemoglu and Restrepo (2019) model automation's labor-market effect as the net result of two counteracting forces: a displacement effect, in which capital replaces labor in tasks it previously performed, and a reinstatement effect, in which new task categories emerge in which labor retains a comparative advantage. Applied to tax administration specifically, this framework predicts that AI-driven automation of routine processing should displace the clerical and data-entry functions that have historically absorbed a substantial share of tax-agency staffing, while reinstating labor into a reconfigured set of functions concentrated around risk-flag review, taxpayer communication, and the discretionary judgement calls automation cannot reliably perform. Autor (2015), revisiting this displacement-reinstatement logic from a complementary angle, argues that automation does not only substitute for labor within a given task category but also raises the productivity, and therefore the value, of the non-routine tasks that remain — an argument with a direct and, this paper suggests, underappreciated implication for tax administration specifically: as AI absorbs routine processing, the discretionary

judgement tax officials continue to exercise over ambiguous or contested cases becomes, if anything, more consequential rather than less, precisely because it is now exercised over the residual category of cases automation could not resolve on its own. This has a direct bearing on the governance concerns developed later in this paper, since it is exactly this residual, judgement-intensive category of case — the kind Robodebt and SyRI both mishandled — where the trust and explainability stakes this paper identifies are highest.

AI and Compliance: From Deterrence to Prediction

Predictive Risk Profiling and Audit Selection

AI's most consequential contribution to compliance specifically concerns risk profiling and audit selection — identifying, from the full population of taxpayers, which returns or entities warrant closer scrutiny. Zheng et al. (2024), surveying data-mining techniques applied to tax risk detection, find machine-learning approaches, including neural networks and ensemble methods, meaningfully improving audit-selection accuracy relative to the threshold-based rules earlier systems relied upon, particularly for detecting complex, multi-entity evasion schemes that simple rule-based screening struggles to identify. This evidence is directly analogous to the fraud-detection evidence documented in the broader accounting literature: Bao, Ke, Li, Yu, and Zhang (2020), developing a machine-learning model for financial-statement fraud detection using raw accounting data, find substantial performance gains over established benchmark models — a methodological pattern Zheng et al.'s tax-specific findings closely parallel.

The Compliance–Trust Paradox

The theoretical tension this paper has developed becomes concrete at exactly this point. Improved audit-selection accuracy, evaluated purely against the deterrence model's logic, should straightforwardly improve compliance by increasing effective detection probability. But Guglyuvatyy (2025) documents that AI-based taxpayer profiling, even where statistically well-calibrated in the aggregate, raises specific fairness concerns when individuals or businesses are flagged as high-risk based on variables — geographic location, business type, historical filing patterns — that may correlate with, without directly measuring, the underlying non-compliance risk the system is meant to detect. Where flagged taxpayers cannot obtain an adequate explanation for their classification, the slippery-slope framework predicts exactly the trust erosion this paper has argued is AI-enabled compliance's central, underexamined risk:

taxpayers who perceive risk profiling as opaque or arbitrary may reduce voluntary compliance specifically in response to that perception, independent of whether the underlying profiling model is, in a narrow statistical sense, accurate.

A related and, this paper argues, equally consequential risk concerns how tax officials themselves respond to AI-generated risk flags once those flags reach the human review stage the governance frameworks reviewed later in this paper generally require. Kim, Glaeser, Hillis, Kominers, and Luca (2024), evaluating a pilot comparing algorithms of differing sophistication in a municipal inspections context structurally similar to tax audit selection, find that officials retaining override authority exercised it frequently, often without documentation clearly connected to genuine additional information they possessed, and that the resulting decisions were, on average, no better than those made without algorithmic assistance despite the algorithms delivering substantial, measurable prediction gains. Applied to tax administration, this finding carries a direct implication this paper treats as significant: a tax agency can deploy a genuinely more accurate risk-profiling system and see no corresponding improvement in actual audit outcomes, or even a deterioration in outcomes for taxpayers whose cases are affected by inconsistent override behavior, if the officials reviewing algorithmic flags are inclined to accept or dismiss them in ways unconnected to case-specific information the system lacked. This risk compounds, rather than substitutes for, the trust-erosion risk discussed above: a taxpayer whose case was mishandled through inconsistent human override of an accurate algorithmic flag has, from that taxpayer's perspective, experienced exactly the same unexplained, seemingly arbitrary treatment that erodes trust when the algorithm itself is at fault, even though the underlying failure mode is behavioral rather than technical.

AI and Revenue Generation

Beyond compliance narrowly defined, a smaller but growing literature examines AI's broader relationship to tax revenue generation and the institutional conditions under which that relationship holds. Saba and Monkam (2024), examining the interplay among AI adoption, institutional quality, and economic growth across the G-7 economies, find AI's contribution to tax revenue outcomes moderated substantially by institutional quality — AI-enabled tax administration improvements translate into stronger revenue outcomes specifically in institutional contexts with stronger governance and rule-of-law indicators, and considerably less reliably where those institutional conditions are weaker. This finding has direct relevance to this paper's central argument: it suggests the trust and governance conditions this paper has

identified as necessary for AI's compliance benefits to translate into sustainable outcomes are not merely a normative concern about taxpayer fairness, but a measurable determinant of whether AI adoption actually delivers the revenue gains it is often adopted to achieve.

The IRS's own reporting on its AI-modernization initiative, funded substantially through the Inflation Reduction Act, offers a complementary domestic illustration: Commissioner Daniel Werfel credited AI-driven tools with improving return processing and anomaly detection during the 2023 filing season, framing these gains explicitly in terms of both administrative efficiency and taxpayer experience — an implicit acknowledgment that revenue-generation gains from AI adoption are, in the IRS's own account, bundled with taxpayer-facing legitimacy considerations rather than separable from them. This paper reads this bundling as consistent with the slippery-slope logic developed above: revenue authorities that frame AI adoption purely in enforcement terms, without corresponding attention to taxpayer experience and trust, risk the compliance-trust paradox this paper has identified, while authorities that frame AI adoption as simultaneously improving efficiency and taxpayer experience are, whether deliberately or not, addressing both dimensions the slippery-slope framework identifies as jointly necessary for sustainable compliance.

Challenges Facing Modern AI-Enabled Tax Systems

Algorithmic Bias and the Robodebt and SyRI Lessons

The two cases this paper has referenced throughout warrant direct examination as theoretically generative failures rather than isolated scandals. Australia's Robodebt scheme automated the identification of welfare overpayments by averaging annual income data across fortnightly reporting periods, a methodology that systematically misrepresented income for anyone with variable earnings and improperly reversed the burden of proof, requiring individuals to disprove automatically generated debt notices rather than requiring the system to establish the debt's validity in the first instance. The Royal Commission into the scheme found it unlawful, and its downstream costs — compensation payments, reputational damage, documented psychological harm to affected individuals — substantially exceeded whatever administrative savings the automation was intended to produce. The Netherlands' System Risk Indication program was found by the District Court of The Hague to violate the right to privacy under Article 8(2) of the European Convention on Human Rights, specifically because the risk-profiling methodology could not be adequately explained even to the court reviewing it, and because the system had relied in part on nationality as a risk factor without an adequate lawful basis — a finding that drew explicit condemnation from the UN Special Rapporteur on extreme

poverty and human rights for its potential to entrench existing social inequality through ostensibly neutral algorithmic means.

Both cases illustrate the same underlying theoretical point from different angles: Kellogg, Valentine, and Christin (2020), examining algorithmic control more broadly, find that AI-based systems function simultaneously as tools that support decision-making and as mechanisms through which organizations, including public administrations, exercise control over the individuals those systems assess — a duality that becomes especially consequential when the individuals being controlled are citizens with legally protected due-process rights rather than employees within a private organizational hierarchy. Munoko, Brown-Liburd, and Vasarhelyi (2020), examining the ethical implications of AI in auditing, identify accountability for AI-generated conclusions as a distinct organizational risk — a risk that Robodebt and SyRI both illustrate at the level of an entire government agency rather than an individual firm, with correspondingly larger consequences when accountability structures prove inadequate.

Explainability and the Right to Reasons

Both cases also share a specific procedural failure this paper treats as theoretically central: neither system could adequately explain individual determinations to the people those determinations affected, a failure the taxpayer-rights literature Guglyuvatyy (2025) reviews frames as a violation of the well-established “right to be informed” embedded in most jurisdictions’ taxpayer charters and administrative-law traditions. Kokina and Davenport (2017), tracing the early stages of AI adoption in accounting and assurance work more broadly, note that automation was beginning to change professional work considerably before generative AI or the more sophisticated risk-profiling systems examined in this paper existed — a historical point this paper reads as evidence that the explainability problem is not confined to the most advanced AI systems currently in use, but has been a latent risk throughout tax administration’s longer digitalization trajectory, becoming visible specifically where automated systems began making or materially influencing individual determinations rather than merely processing data.

Cross-Border Coordination and Regulatory Fragmentation

A further challenge concerns the fragmented state of AI-specific regulation across the jurisdictions in which multinational taxpayers and, increasingly, AI service providers themselves operate. The European Union’s AI Act draws a formal distinction between general-

purpose and high-risk AI systems and imposes transparency obligations accordingly, but Guglyuvatyy (2025) finds that even this relatively developed regulatory framework does not directly address AI's role in tax administration specifically, leaving what this paper treats as a significant regulatory gap at the intersection of automated governance and taxpayer rights. The National Institute of Standards and Technology's AI Risk Management Framework (NIST, 2023) offers an instructive point of comparison: unlike the sector-specific and still-developing tax regulations reviewed above, the NIST framework organizes AI governance around four functions — govern, map, measure, and manage — that explicitly treat accountability and transparency as first-order design requirements to be established before an AI system is deployed, rather than as considerations addressed only after deployment reveals a problem. Tax administrations that have adopted the NIST framework, or comparable govern-first sequencing, alongside their existing tax-specific compliance obligations are, on this paper's reading, considerably better positioned to avoid the retrofitted-explainability problem that characterized both Robodebt and SyRI than administrations relying solely on generic public-sector procurement and privacy requirements not designed with AI-specific risks in mind.

Dwivedi et al. (2023), addressing generative AI's risks and governance challenges across disciplines more broadly, reach a consistent conclusion from outside the tax context specifically: regulatory frameworks have generally struggled to keep pace with the deployment speed of the AI systems they are meant to govern, a gap with particular consequence in tax administration given the cross-border character of an increasing share of both tax avoidance schemes and the AI systems used to detect them. Zhang, Zhu, Dai, and Chen (2023), examining the ethics of AI in managerial accounting more broadly, identify accountability and transparency as recurring concerns that automation initiatives tend to treat as secondary design considerations rather than central ones — an observation this paper suggests applies with particular force to tax administration specifically, where the consequences of inadequate accountability extend beyond individual organizational harm to encompass the kind of systemic legitimacy damage the Robodebt and SyRI cases both illustrate at national scale.

AI and Taxpayer-Facing Service

Beyond automation, compliance detection, and revenue generation, a distinct and theoretically significant application concerns AI systems designed to serve taxpayers directly rather than to monitor or enforce compliance against them. Natural-language chatbots and virtual assistants, deployed by revenue authorities including the Estonian tax administration's 'Kratt' system and comparable tools piloted elsewhere, allow taxpayers to

pose questions about filing obligations, deductions, and procedural requirements in ordinary language rather than navigating dense statutory text or waiting for a human representative. This application sits apart from the detection-and-enforcement applications discussed earlier in this paper because it operates, at least in principle, on the trust side of the slippery-slope framework directly: a taxpayer who can obtain a clear, immediate, and plain-language answer to a genuine question about their obligations is being served by exactly the kind of transparency and comprehensibility this paper has argued detection-oriented AI applications frequently fail to provide.

This distinction matters theoretically because it suggests AI's relationship to tax-administration trust is not uniformly negative, contrary to what a reading focused solely on Robodebt and SyRI might imply. Guglyuvatyy (2025) argues that AI-based taxpayer-facing systems, if transparently governed and regularly audited, could function as a genuine access-to-justice mechanism, particularly for taxpayers who currently face structural barriers — cost, geographic distance, limited digital literacy — to obtaining professional tax advice or engaging effectively with under-resourced public tax clinics. On this reading, the same underlying AI capability that produces algorithmic bias risk when applied to enforcement can produce genuine trust and access benefits when applied to taxpayer service, which reinforces this paper's central theoretical claim in a specific way: the relevant variable determining whether AI strengthens or weakens the taxpayer-state trust relationship is not the sophistication of the AI technology itself, but the specific function — monitoring the taxpayer or serving the taxpayer — that technology is deployed to perform, and the governance conditions attached to that deployment.

This distinction should not be overstated, however. Taxpayer-facing systems carry their own, structurally different risk this paper's model has to account for: a chatbot or virtual assistant that provides confidently stated but incorrect guidance can mislead taxpayers into non-compliance they did not intend, a risk with some structural resemblance to the fluency-accuracy gap documented in the broader generative-AI literature, where systems capable of producing fluent, plausible-sounding output do not thereby guarantee that output is accurate. A taxpayer-facing system's trust benefits, in other words, are conditional on the same explainability-by-design and governance principles this paper has applied to enforcement-side applications throughout — the function differs, but the underlying governance requirement, on this paper's account, does not.

Bringing the preceding sections together, this paper proposes that AI's effect on tax

administration be evaluated through three parallel outcomes — automation efficiency, compliance behavior, and revenue generation — each grounded in the theoretical mechanisms reviewed above, and each conditioned by a governance layer addressing explainability, human oversight, and taxpayer redress. Figure 1 summarizes this model. AI capabilities feed into the compliance-theory mechanisms this paper has reviewed — deterrence, trust and power, and the principal-agent relationship — which in turn shape compliance behavior and revenue outcomes; but both outcomes pass through a governance layer this paper's evidence suggests is currently underdeveloped relative to the detection and processing capability driving it, with the Robodebt and SyRI cases illustrating what happens when that governance layer proves inadequate to the automated decision-making it is meant to support.

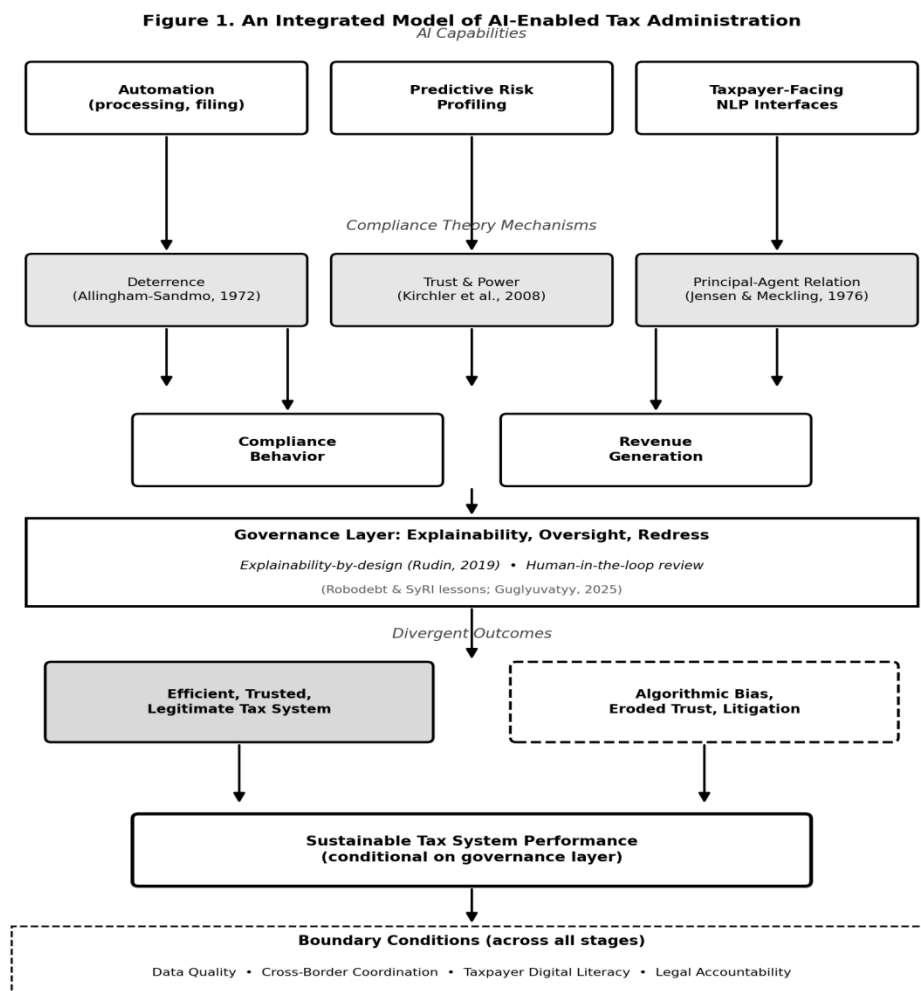


Figure 1. An Integrated Model of AI-Enabled Tax Administration, showing AI capabilities operating through compliance-theory mechanisms toward divergent system outcomes, conditioned by a governance layer addressing explainability, oversight, and redress.

This model's central implication is that the two facts this paper opened with —

substantial detection gains and consequential legitimacy failures — are not competing narratives about AI in tax administration but complementary evidence for the same underlying structure: detection and processing capability have advanced considerably faster than the governance layer needed to convert that capability into sustainable, trust-preserving compliance outcomes. Saba and Monkam's (2024) finding that AI's revenue benefits are institutionally conditional supports this reading directly — the same technology that improves outcomes in strong-governance contexts can fail to do so, or actively undermine trust, where governance conditions are weaker, which is precisely the pattern this paper's model predicts.

Implications for Tax Policy and Administration

For tax administrations, the model developed here suggests explainability should be treated as a design requirement from the outset of AI system development, consistent with Rudin's (2019) interpretability-by-design principle, rather than as a compliance obligation addressed after a system's opacity has already generated legal or reputational harm. The Robodebt and SyRI cases both suggest that retrofitting explainability onto an already-deployed opaque system is a considerably more costly path than designing for it from the beginning, both in direct remediation costs and in the harder-to-reverse trust damage the slippery-slope framework identifies as tax administration's most consequential long-run risk.

For audit-selection and risk-profiling systems specifically, this paper's evidence suggests administrations should track not only detection accuracy but taxpayer-perceived fairness and comprehensibility as separate, explicitly monitored outcomes, since Guglyuvatyy's (2025) review indicates these can diverge substantially even where aggregate statistical performance is strong. Human-in-the-loop review, in the sense the broader algorithmic-oversight literature recommends, should be structured specifically to preserve tax officials' genuine capacity to interpret and, where warranted, override algorithmic determinations — a capacity the Australian National Audit Office found had measurably eroded among Australian Taxation Office staff even where formal human-oversight requirements remained technically in place.

For policymakers and regulators, the gap this paper has identified between general AI regulation, such as the EU AI Act, and tax-administration-specific application suggests a need for domain-specific guidance that addresses the particular due-process and burden-of-proof considerations tax and welfare determinations carry, rather than relying on general-purpose AI governance frameworks to address these considerations by extension. Given the cross-border

character of both sophisticated tax avoidance and the AI systems increasingly used to detect it, this paper's evidence also points toward a need for international coordination on tax-specific AI governance standards comparable to the coordination that already exists around information exchange and base-erosion prevention.

A further implication concerns how tax administrations communicate AI adoption to the taxpaying public more broadly, independent of any individual taxpayer's direct interaction with a specific algorithmic determination. The IRS's own framing of its AI-modernization program, bundling efficiency gains with taxpayer-experience improvements rather than presenting them as separate objectives, offers one template consistent with this paper's model; administrations that instead frame AI adoption primarily or exclusively in enforcement terms — emphasizing detection capability and compliance-gap closure without corresponding attention to taxpayer-facing transparency — risk signaling exactly the power-without-trust configuration the slippery-slope framework predicts should shift compliance toward its more costly, enforcement-dependent equilibrium. This paper's reading of the evidence suggests communication strategy is not a peripheral concern relative to the underlying technology, but a direct lever on which of the two divergent outcomes this paper's model identifies actually materializes.

Directions for Future Research

Several gaps follow from the model developed above. First, the compliance-trust paradox this paper has theorized — that AI-enabled detection can simultaneously improve power-based deterrence and erode trust-based voluntary compliance — has not been directly tested using the slippery-slope framework's own established measurement instruments; research applying validated trust-and-power surveys to taxpayers in jurisdictions before and after significant AI-enabled compliance system deployment would test this paper's central theoretical claim far more directly than the case-study evidence currently available.

Second, Saba and Monkam's (2024) finding that institutional quality moderates AI's revenue effects has been established at the level of G-7 economies specifically; extending this analysis to a broader and more institutionally heterogeneous sample of countries, including lower-income and weaker-governance contexts where AI-enabled tax administration is increasingly being piloted, would test whether the institutional-conditionality this paper's model predicts holds across a wider range of governance environments.

Third, the explainability-by-design principle this paper has drawn from Rudin (2019) and

applied to tax administration has not been empirically tested in this specific domain; research comparing taxpayer trust and dispute-resolution outcomes between administrations using inherently interpretable risk-profiling models and those using more complex but less interpretable alternatives would give this paper's central practical recommendation an empirical foundation it currently lacks.

Finally, comparative research examining how the Robodebt and SyRI governance failures have, or have not, changed subsequent AI-system design and deployment practices within the agencies involved, and within other tax administrations that have studied these cases, would clarify whether high-profile legal reversals of this kind actually produce durable governance learning, or whether the underlying structural gap this paper has identified between detection capability and governance capacity tends to reassert itself as AI capability continues to advance.

A sixth and final gap concerns the taxpayer-facing services this paper has distinguished from enforcement-oriented applications; the access-to-justice potential Guglyuvatyy (2025) and this paper both attribute to well-governed taxpayer-facing AI systems remains largely theoretical, drawn from a small number of pilot deployments such as Estonia's Kratt system rather than from systematic evaluation. Research directly comparing taxpayer comprehension, satisfaction, and downstream compliance behavior between jurisdictions that have deployed taxpayer-facing conversational AI at scale and those that have not would test whether this paper's proposed trust-building function of taxpayer-facing AI actually materializes in practice, and would usefully complement the considerably more developed literature on AI's enforcement-side effects this paper has drawn on throughout.

Conclusion

Artificial intelligence is demonstrably strengthening tax administrations' capacity to process returns efficiently and detect non-compliance accurately, and the evidence reviewed in this paper — from machine-learning-based audit selection to large-scale production deployment at major revenue authorities — leaves little doubt that these gains are real and continuing to grow. This paper has argued, however, that evaluating AI's contribution to tax administration on these terms alone significantly understates what is actually at stake, because sustainable tax administration depends on a trust relationship between taxpayer and state that classical deterrence theory, the slippery-slope framework, and agency theory all identify as necessary for efficient, low-cost, predominantly voluntary compliance — and AI-enabled

enforcement, however statistically accurate, does not automatically preserve that relationship. The Robodebt and SyRI cases demonstrate concretely what this paper has argued theoretically: a tax administration can become more accurate in the aggregate while becoming less legitimate, less explicable, and ultimately less trusted in the particular, with consequences that extend well beyond the individual taxpayers directly affected.

None of this argues against continued AI investment in tax administration; the efficiency and detection evidence reviewed throughout this paper is unambiguous that such investment produces genuine value, and the alternative of forgoing AI-enabled analytics entirely would leave modern tax administrations poorly equipped to address the scale and sophistication of contemporary non-compliance and avoidance. The argument, rather, is that the value of that investment cannot be assessed by detection accuracy alone, and that administrations confident in their algorithmic capability while under-investing in the explainability, oversight, and redress mechanisms this paper's model identifies risk discovering, as Australia and the Netherlands both did, that the discovery arrives through litigation and public scandal rather than through their own governance processes.

The practical and theoretical implication is the same: tax administrations, policymakers, and researchers should stop evaluating AI's contribution to tax systems primarily by asking whether detection accuracy has improved, and start asking the more demanding question this paper's model makes explicit — **whether the governance layer surrounding that detection capability preserves taxpayers' capacity to understand, contest, and trust the determinations AI increasingly helps produce**. Modern tax systems' revenue-generating capacity has always depended on this trust as much as on enforcement capability; the evidence reviewed in this paper suggests AI is improving the latter considerably faster than it is preserving the former, and closing that gap deliberately is the central challenge this paper's synthesis leaves for tax administration in the years ahead.

References

- Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2), 3–30.
<https://doi.org/10.1257/jep.33.2.3>
- Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3–4), 323–338. [https://doi.org/10.1016/0047-2727\(72\)90010-2](https://doi.org/10.1016/0047-2727(72)90010-2)
- Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3–30.
<https://doi.org/10.1257/jep.29.3.3>
- Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, 58(1), 199–235. <https://doi.org/10.1111/1475-679X.12292>
- Belahouaoui, R., & Attak, E. H. (2024). Digital taxation, artificial intelligence and Tax Administration 3.0: Improving tax compliance behavior—a systematic literature review using textometry (2016–2023). *Accounting Research Journal*, 37(2), 172–191.
<https://doi.org/10.1108/ARJ-12-2023-0372>
- Bruckner, C., & Coil, C. (2024). *AI and the modern tax agency: Adopting and deploying AI to improve tax administration*. American University Kogod Tax Policy Center; IBM Center for the Business of Government.
- Choi, J. H., & Xie, C. L. (2026). Human + AI in accounting: Early evidence from the field. *Journal of Accounting Research*, 64(3), 1333–1373. <https://doi.org/10.1111/1475-679X.70052>
- Cooper, L. A., Holderness, D. K., Sorensen, T. L., & Wood, D. A. (2019). Robotic process automation in public accounting. *Accounting Horizons*, 33(4), 15–35.
<https://doi.org/10.2308/acch-52466>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, Article 102642.

<https://doi.org/10.1016/j.ijinfomgt.2023.102642>

Guglyuvatyy, E. (2025). Balancing innovation and integrity: AI in tax administration and taxpayer rights. *Humanities and Social Sciences Communications*, 12, Article 1818.

<https://doi.org/10.1057/s41599-025-06099-7>

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360.

[https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1), 366–410.

<https://doi.org/10.5465/annals.2018.0174>

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2024). Decision authority and the returns to algorithms. *Strategic Management Journal*, 45(4), 619–648.

<https://doi.org/10.1002/smj.3569>

Kirchler, E., Hoelzl, E., & Wahl, I. (2008). Enforced versus voluntary tax compliance: The “slippery slope” framework. *Journal of Economic Psychology*, 29(2), 210–225.

<https://doi.org/10.1016/j.joep.2007.05.004>

Kokina, J., & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. *Journal of Emerging Technologies in Accounting*, 14(1), 115–122. <https://doi.org/10.2308/jeta-51730>

McCarty, L. T. (1977). Reflections on TAXMAN: An experiment in artificial intelligence and legal reasoning. *Harvard Law Review*, 90(5), 837–893.

<https://doi.org/10.2307/1340132>

Munoko, I., Brown-Liburd, H. L., & Vasarhelyi, M. (2020). The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics*, 167(2), 209–234.

<https://doi.org/10.1007/s10551-019-04407-1>

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce.

<https://doi.org/10.6028/NIST.AI.100-1>

Organisation for Economic Co-operation and Development. (2025). *Tax administration digitalisation and digital transformation initiatives*. OECD Publishing.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

- Saba, C. S., & Monkam, N. F. (2024). Leveraging the potential of artificial intelligence (AI) in exploring the interplay among tax revenue, institutional quality, and economic growth in the G-7 countries. *AI & Society*. Advance online publication. <https://doi.org/10.1007/s00146-024-01885-4>
- Sutton, S. G., Holt, M., & Arnold, V. (2016). “The reports of my death are greatly exaggerated”—Artificial intelligence research in accounting. *International Journal of Accounting Information Systems*, 22, 60–73. <https://doi.org/10.1016/j.accinf.2016.07.005>
- Zhang, C., Zhu, W., Dai, J., & Chen, X. (2023). Ethical impact of artificial intelligence in managerial accounting. *International Journal of Accounting Information Systems*, 49, Article 100619. <https://doi.org/10.1016/j.accinf.2023.100619>
- Zheng, Q., Xu, Y., Liu, H., Shi, B., Wang, J., & Dong, B. (2024). A survey of tax risk detection using data mining techniques. *Engineering*, 34, 43–59. <https://doi.org/10.1016/j.eng.2023.07.014>