Information Systems Journal

WILEY

SPECIAL ISSUE ARTICLE OPEN ACCESS

How to Demonstrate Trustworthy Use of AI in Public Services: A Case Study

Natalie Smith[1] | Nicole Gillespie[2,3] | Tapani Rinta-Kahila[1,4] | Steven Lockey[1] | Javad Pool[5] | Caitlin Curtis[1,6]

[1]Business School, The University of Queensland, Brisbane, Queensland, Australia | [2]Melbourne Business School, The University of Melbourne, Melbourne, Victoria, Australia | [3]Centre for Corporate Reputation, The University of Oxford, England, United Kingdom | [4]Hanken School of Economics, Helsinki, Finland | [5]Centre for Work, Organisation and Wellbeing, Griffith University, Brisbane, Australia | [6]Centre for Policy Futures and Business School, The University of Queensland, Brisbane, Queensland, Australia

Correspondence: Natalie Smith (natalie.smith@business.uq.edu.au)

Received: 11 April 2024 | **Revised:** 25 September 2025 | **Accepted:** 14 October 2025

Keywords: artificial intelligence | public sector | trust | trustworthy AI

ABSTRACT

Government leaders across the globe are grappling with how to harness and integrate artificial intelligence (AI) to enhance public service delivery and efficiency. Yet, a key challenge faced is how to build and maintain the trust of stakeholders. Trust is critical for the acceptance and sustained adoption of AI technologies, as well as to gain the requisite funding, resourcing and authorization to implement AI solutions. However, inherent features of AI—its autonomous capabilities, dynamic learning, and inscrutable operating logic—create challenges for trust, particularly in public services that are subject to high expectations of accountability, transparency, and fairness. We present an in-depth case analysis of how an Australian government department was able to deploy a solution that was widely accepted, and identified as an exemplar of trustworthy AI use. We identify six trust-supporting approaches: benevolent customer-centricity, radical honesty, diverse input, rigorous development and testing, human discretion in decision-making, and aligning the authorising environment. For each approach, we explain how and why it supports trust, and then contrast that approach with a prominent, but widely distrusted application in the Australian government. We conclude with implications for public sector leaders seeking to engender trust in their use of AI.

1 | Introduction

Artificial intelligence (AI) is fueling a revolution in the administration and governance of public services due to its potential to increase efficiency, overcome resource constraints, personalise services, and address long-standing societal issues (Manzoni et al. 2022; Medaglia et al. 2023). However, public trust in governments' use of AI has been called into question, contributing to slow uptake, acceptance, and adoption of AI in public services (Gillespie et al. 2021, 2023; Margetts and Dorobantu 2019).

For government leaders to engender trust in this technology is far from straightforward. Citizens' trust in public authorities is critical due to their social licence to operate, yet trust in government is declining (Pew Research Centre 2024). Governments hold many forms of authoritative power over citizens, so public scrutiny and expectations of accountability, transparency, fairness and privacy in the public sector are high (Andrews et al. 2022; Kalesnikaite and Baker 2025; Ring and Perry 1985). This power stems from the government's distinctive responsibility for administrative law—the body of law regulating government decision-making—and the structural separation of the legislature, executive and judiciary. This gives the government legitimate authority over many aspects of an individual's life, including the right to tax, penalise, and incarcerate, and have privileged access to sensitive information about citizens (Attorney-General's Department 2011; NSW Ombudsman 2021b; Tyler 2006). Upholding procedural norms

and values of impartiality, accountability, and transparency in the exercise of power—including discretionary decision-making and the delivery of public services—is central to maintaining government legitimacy and trust (Kalesnikaite and Baker 2025; Rothstein and Teorell 2008; Tyler 2006; Weber 1978).

AI systems that are based on machine learning (ML) models are characterised by three interdependent facets—autonomy, learning and inscrutability—differing fundamentally from previous generations of organisational technologies (Berente et al. 2021). These unique characteristics of AI come into direct conflict with the expectations people assign to the public sector, making it difficult to implement models in a trusted and trustworthy manner (Desouza and Dawson 2023).

First, accountability for decisions that affect citizens is a critical requirement for public organisations (Manzoni et al. 2022). Yet the unprecedented ability of AI to operate autonomously can make determining accountability for decisions complicated and unclear (Berente et al. 2021). This is particularly problematic in high-stakes public service contexts, such as determining parole decisions (Dwivedi et al. 2021; Margetts and Dorobantu 2019). Second, public sector decisions are expected to be consistent, fair, and defensible (NSW Ombudsman 2021b), which is challenged by the dynamic, self-learning capability of AI systems, particularly AI models that automatically learn based on incoming data after their deployment (Russell and Norvig 2016). This continuous adaptation can lead to unpredictable outcomes, making it difficult to ensure that decision-making based on such outcomes is fair and defensible. Third, due to these dynamic learning processes, AI models tend to be complex, and their operating logic can often be inscrutable to their users (Asatiani et al. 2021; Faraj et al. 2018). Such inscrutability can be unacceptable in the public sector where transparency of decisions is often expected. Furthermore, AI models can require vast amounts of data on citizens, amplifying concerns over privacy and power imbalance.

Compounding the challenge is that the governance and assurance frameworks in government—which normally provide a level of confidence and safety to mitigate risks—are still evolving for AI (for e.g., Department of Finance 2024; European Commission 2022; GAO 2021; Office for Artificial Intelligence 2022). Government has a key role in developing these regulations and creating the right environment for responsible AI (DTA 2024), and there is a risk to integrity and the social licence to regulate and govern others if governments are not acting responsibly themselves. These tensions raise challenges for public sector leaders who are increasingly expected to leverage the benefits of AI to enhance service delivery and efficiency and reduce costs (van Noordt et al. 2023).

While recent research has tackled trust-related challenges such as inscrutability (Asatiani et al. 2020; Someh et al. 2022) and the shift of agency from humans to AI (Shollo et al. 2022), critical questions remain around how to overcome public scepticism and distrust of government AI use, and how to secure the trust of key stakeholders to enable investment and deployment of AI projects (Margetts and Dorobantu 2019; van Noordt et al. 2023). Consequently, our research asked: *how can government leaders demonstrate to stakeholders that their use of AI in public services is trustworthy?* We consider two main stakeholder groups: internal government employees who develop, operate, govern, have accountability for, or provide resources for the AI solution, and external stakeholders who regulate, influence, rely on, or are subjected to the AI-enabled services.

To answer this question, we report on a case analysis of an AI solution developed and implemented in an Australian state government revenue department. We draw on in-depth interviews with those responsible for developing, deploying, operating and governing the solution and those who represent the interests of affected citizens, triangulated with case documentation.

The case provides a particularly rich and informative context for studying how trust challenges can be navigated for several reasons. First, the case focuses on the application of AI in government debt collection from citizens, a context characterised by high distrust from multiple stakeholders (Henley 2021; Lindebaum et al. 2023; Van Bekkum and Borgesius 2021). Second, the AI solution became widely accepted despite being implemented soon after a high-profile technology failure that eroded public trust in the Australian government's use of technology (see Lindebaum et al. 2023; Rinta-Kahila et al. 2022). Third, the case occurred at a time when there were limited established policies or frameworks to guide the trustworthy implementation of AI. Indeed, this case became an exemplar that informed the subsequent formation and institutionalisation of responsible AI frameworks in the Australian public sector, helping other early AI adopters and innovators in government (NSW Government 2021a). The ability to build and sustain acceptance in this challenging context underscores the robustness of the approaches taken to demonstrate trustworthiness and support trust.

2 | Defining AI, Trust and Trustworthiness

To clarify what we mean by trust in AI solutions, we first define the terms and identify key factors that influence trust. We use the OECD definition of an *AI solution* as 'a machine-based system that for explicit or implicit objectives, infers, from the inputs it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments' (OECD.AI 2023).

Trust is commonly defined as a person's willingness to be vulnerable to the actions of another party or entity, based on positive expectations of the future intentions and actions of that party (Mayer et al. 1995; Rousseau et al. 1998). Trust is a key determinant of sustained AI acceptance and adoption (Kelly et al. 2023; Misra et al. 2023), broader citizen compliance and cooperation with public authorities (Tyler 1998), and public policy effectiveness (Lundin 2007). While trust is a complex, contextual and dynamic process (Gillespie and Dirks 2026), a central proposition to theories of trust is that the willingness to be vulnerable to an entity is grounded in 'good reasons' (Lewis and Weigert 1985; Mollering 2006, 13).

One important source of 'good reasons' to trust comes from the assessment of the entity's *trustworthiness*. Literature has identified three key dimensions of trustworthiness—ability, benevolence, and integrity—which capture a set of key factors influencing trust (Colquitt and Rodell 2011; Mayer et al. 1995), including

trust in public sector agencies (Devine et al. 2025; Gillespie and Daly 2025) and technology, and AI implementations (Bedué and Fritzsche 2022; Laux et al. 2024). *Ability* refers to government competence and capability to effectively deliver on services and responsibilities (Devine et al. 2025; Mansoor 2021; van Erkel and van Der Meer 2016). In the context of AI, this includes ensuring AI systems perform reliably to produce accurate output as expected. *Benevolence* refers to the perception that government cares about citizens' well-being, interests and needs, and operates fairly in the best interest of the public (Devine et al. 2025; Mayer et al. 1995). This includes procedural fairness and consistency in decision making (Grimes 2006), providing citizens a voice and participation in government processes (Prats et al. 2023), and in the context of AI, ensuring AI systems result in positive outcomes and minimise harm. *Integrity* refers to government operating honestly, transparently, and with accountability (e.g., Birch and Allen 2010; Ouattara et al. 2023), including following through on commitments and ensuring AI-enabled services uphold ethical principles, values, and laws.

Another source of 'good reasons' to trust comes from structural assurances and control mechanisms, which can take the form of regulation, laws, rules, standards, governance, and oversight mechanisms. For any technology, these mechanisms provide an important basis for trust (McKnight et al. 2011). In the context of AI, they help increase trust by clarifying expected standards and practices, incentivising adherence to rules and standards, and constraining untrustworthy use (e.g., through disincentives and punitive consequences for rule-breaking; Long and Sitkin 2018; Park 2020; Shao et al. 2020).

3 | Focus of Our Research

We focus our analysis predominately on how government leaders demonstrate *trustworthiness* in their use of AI as this is what 'earns' sustained, well-placed trust over time. It is the demonstration of trustworthiness—rather than trust itself—that is within the agency and control of organisational actors and provides the foundation for establishing trust (O'Neill 2018). In contrast, whether or not a stakeholder gives trust in any given situation is at the discretion of that individual stakeholder.

For this study, we take a *sociotechnical stance*, viewing trust in an AI-enabled service as informed by a combination of the trustworthiness of the AI system and the organisational and regulatory context within which it is embedded (Söllner et al. 2016; Van der Werff et al. 2021). Specifically, this involves consideration of the technical *AI system* (e.g., its functionality, reliability, predictability, and helpfulness), the *deployment and use* of that system (e.g., is it being implemented and used appropriately; will it benefit stakeholders?), the *organisation* deploying, using, and governing the system (e.g., does the organisation operate with integrity and have benevolent intentions?), and the broader regulatory and legal *institutional context*.

4 | The Revenue NSW AI Case

To bring insight into how government leaders can support stakeholders' trust in the integration of AI into service delivery, we conducted a case study of an AI solution developed and implemented by an Australian public sector agency, Revenue NSW. This case was one of eight case studies involving over 100 interviews conducted as part of a larger examination of trust in AI-enabled government services.[1] We selected this case because the solution was well received, widely accepted and identified by central government agencies as an exemplar of trustworthy AI.

Our case analysis is based on interviews triangulated with case documentation. In-depth interviews were conducted with stakeholders responsible for developing, deploying, leading, using and governing the AI solution (e.g., developers, domain experts, project leads, agency leaders, the responsible minister, and external regulators). Case documentation included media articles, parliamentary transcripts, independent reports, Department presentations, and public communications about the AI solution on agency websites and social media (see Appendix A for further details of data collection and method rationale).

4.1 | Case Context

Revenue NSW collects revenue owed to the government by individuals and organisations within the government's jurisdiction in the state of New South Wales (NSW). It collects over AU$40 billion per annum on behalf of over 240 government-related entities from a combination of payroll tax, duties (e.g., transferring property ownership), fines (e.g., traffic infringements), fees (e.g., licence fees), and royalties (e.g., related to mining) (DCS 2022). The Department has over four million customer records, with approximately 60 000 records associated with people considered to be vulnerable due to their susceptibility or exposure to physical, emotional, social, or financial harm (NSW Ombudsman 2021a). The Department has the right to collect significantly overdue unpaid debts via garnisheeing—a practice of taking the owed money directly from the debtor's bank account. However, this practice is highly problematic when applied to vulnerable people.

Originally, Revenue NSW had manually and inconsistently administered garnisheeing, limited by the number of agency employees relative to the thousands of accounts in debt. To improve efficiency and consistency, the Department automated the garnisheeing of accounts, which increased from 6905 transactions per annum in 2010 to 1.6 million in 2018 (NSW Ombudsman 2021a). Although only 2% of these garnishee orders were successful (Revenue NSW 2021b), the automated process did not consider the debtors' vulnerability, which led to a surge of complaints and caused an increasing work burden on the Department. The complaints brought the Department under increasing public scrutiny with negative media attention and an inquiry from the Ombudsman, who questioned the legality of the Department's automatic garnisheeing practices (NSW Ombudsman 2021a, 2021b). The inquiry determined that while the authority to garnishee was discretionary and legal, full automation of garnisheeing that removed human discretion and judgement from the process was unlawful.

In response, the Department set out to transform its 'very gung-ho' and 'compliance-focused' (P70) debt collection process to a more customer-focused approach. The overall program

instituted alternatives to repaying debt for vulnerable people, such as 'education programs, drug rehabilitation programs, psychology, and education-type programs' (P19), with the intent to achieve better long-term individual and community outcomes. To achieve this goal, it was necessary to identify customers who were vulnerable and/or unable to pay, bring empathy to service provision, and find alternative approaches for these customers' individual circumstances. Given the sheer number of unpaid accounts, and the volume of data that was needed, AI was considered a potential solution to help identify vulnerable customers and distinguish them from customers who could legitimately pay:

> What I would say was really good in [Revenue NSW] is the level of information and data that we had on what was happening. Because our processes were heavily digitalized, we knew at all times how many fines had been issued, [...] how many people are paying an enforcement order versus not. And then the real missing piece on that was – well, who's not paying because they can't pay and who's not paying just because they don't believe in the system and just don't want to pay fines.
>
> (P57)

4.2 | The AI Solution

The AI solution was introduced in 2018 to identify individuals likely to be vulnerable and prevent automatic garnisheeing of their accounts. The solution identifies vulnerability based on over 20 attributes, such as the debtor's age, address, severity of fines, the amount owed, type of offence, and known incarceration history (Revenue NSW 2021a, 2021b). The solution's output is a relative measure of a customer's likeness to typical vulnerability characteristics:

> Essentially what the model is doing...it's taking those customers that are known to be vulnerable, building likeness characteristics, and then saying, "Well, what's the likelihood of the other customers [being vulnerable]" and ... giving us basically a prediction of how closely...do they match a person that we know to be vulnerable.
>
> (P51)

A benevolent approach was taken where cases flagged as vulnerable by the AI system are then reviewed by a case worker who contacts the person to assess vulnerability and decide whether or not to negotiate non-financial alternatives. The solution was developed in-house using operational funds with publicly accessible open-source tools and consequently was relatively low-cost.[2] The AI-enabled service was supported by introducing a protected 'minimum balance' for accounts being garnisheed in 2016, a benevolent approach designed to protect economically vulnerable people, and implementing human oversight of the process for garnisheeing accounts in 2019.

4.3 | An Exemplar of Trustworthy Public Sector AI Use

The overall program of offering alternatives to debt collection for vulnerable people—enabled by the AI solution—won an across-government award in 2019 for successfully tackling long-standing social challenges (Revenue NSW 2020). NSW Government reported that 'with approximately 46 000 customers considered "vulnerable", the community benefits of an effective AI solution are substantial [and] results in fewer vulnerable people being forced to pay fines that they cannot afford' (NSW Government 2021a).

In 2021, the AI solution was identified as an exemplar of trustworthy public sector AI use by the Government (NSW Government 2021a). The case was retrospectively 'stress tested', independently verified by peer review and confirmed by the AI Review Committee (NSW Government 2021a) It was found to align with the Government's AI Ethics policy and the key ethical principles: Community Benefit, Fairness, Privacy and Security, Transparency and Accountability. The case was described as demonstrating 'how the NSW Government is developing AI responsibly with a clear focus on outcomes so that the community can have trust that the technology is being used appropriately' (NSW Government 2021a). A timeline of the evolution of the solution is provided in Figure 1.

4.4 | Trust Challenges at Revenue NSW

While Revenue NSW's AI solution was considered an exemplar, and ultimately resulted in sustained stakeholder support, acceptance and adoption, it required navigating multiple thorny challenges.

To begin with, the Department's automated garnisheeing practices attracted negative media attention and the scrutiny of the Ombudsman, which resulted in a general distrust of its ability to responsibly manage technology implementations. This resulted in a report written by the NSW Ombudsman (2021b) that stated:

> We were prompted to write this report after becoming aware of one agency (Revenue NSW) using machine technology for the performance of a discretionary statutory function (the garnisheeing of unpaid fine debts from individuals' bank accounts), in a way that was having a significant impact on individuals, many of whom were already in situations of financial vulnerability.

As one participant stated, this criticism led to a 'degree of sensitivity' as 'it's quite hard for any public sector organization to receive the criticism, particularly from somebody as vital as the Ombudsman' (P36).

This distrust was fueled by a high-profile failure of an automated decision-making system by the federal government agency Centrelink (now Services Australia). This system, dubbed 'Robodebt', automatically issued debt notices (sometimes

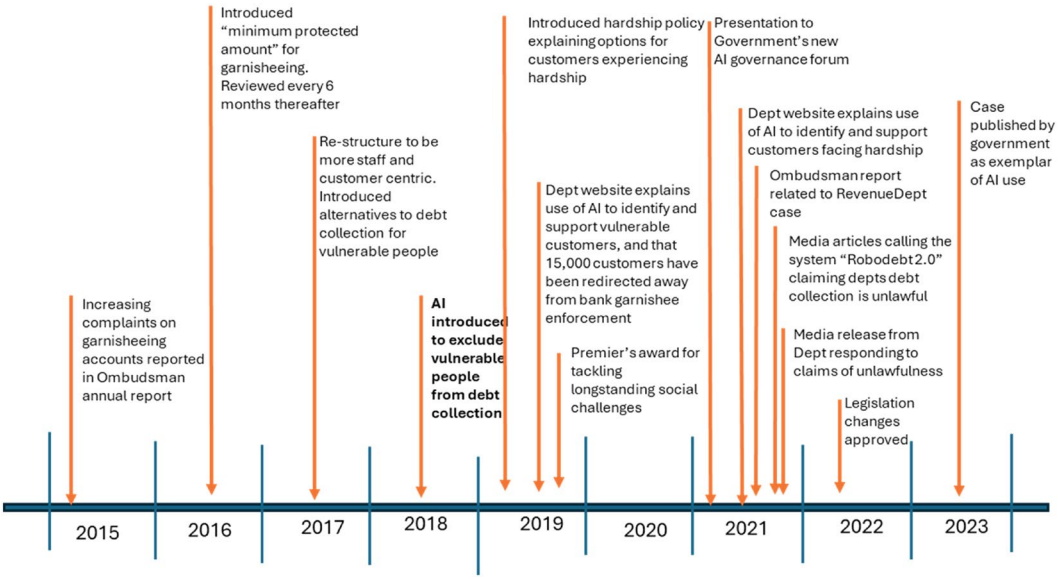

FIGURE 1 | Evolution of revenue NSW's AI solution.

erroneously) to vulnerable welfare recipients, causing widespread distress among citizens, and was ultimately deemed unlawful (Rinta-Kahila et al. 2022). The Australian public's trust in their government's ability to implement technologies responsibly was significantly damaged by Robodebt, along with media scandals and legal proceedings that followed its implementation (Holmes 2023; Rinta-Kahila et al. 2022). Government use of automation to claim payments from citizens became a highly inflammatory topic.[3] Addressing this challenge was not straightforward as it needed to be balanced with the charter and responsibility of Revenue NSW to collect debt on behalf of other government agencies and ensure appropriate revenue generation on behalf of the government.

On top of this, Revenue NSW was one of the early adopters of AI in the public sector, and legislative frameworks that would typically guide such implementations were lagging behind practice. The nascent state of AI in government was also apparent in the area of data sharing. While access to data from other agencies could improve the accuracy of the AI's prediction of vulnerability, data governance protocols were not yet sufficiently mature to facilitate data sharing. As a result, there were concerns about the sufficiency of data points to robustly train the AI. The Robodebt failure compounded this concern, serving as a warning of what can happen when data from different government departments are combined haphazardly.

Finally, there were more general concerns about AI's ability to accurately determine vulnerability. Vulnerability is a dynamic state that can change over time and is typically informed by a combination of factors, including information that may never be visible to the Department. Vulnerability is understood differently across various fields of research, making it a nuanced and challenging concept to define (Hamm et al. 2024). Could a machine really identify vulnerability? How could codifying such contested knowledge into an AI solution be accurate? The Department could not have all the information and history on a particular case due to practical limitations in collecting relevant data, the limited contact vulnerable people often have with government departments, and challenges in gaining consent from individuals, even if the ability to share data across government departments could be resolved.

4.5 | The Stakeholder Context

There were various stakeholders whose trust was required or pertinent for the AI solution to be developed, deployed, operated, and accepted. We note that there is not always a clear delineation between the trustor (i.e., the person doing the trusting) and trustee (i.e., the object of trust) within complex socio-technical systems. Actors can simultaneously be both trustors and trustees within such systems (Lewis and Weigert 1985). For example, the executive in charge of the AI solution is both a trustor (of the teams who build and implement the solution) and a trustee (of the responsible minister to effectively govern the solution). Similarly, internal customer service employees are both trustors (of the AI system) and trustees (of the customers they serve). Table 1 lists these stakeholders and key reasons why their trust in Revenue NSW's AI solution was important.

Internal stakeholders included those who held accountability for the AI-enabled service, such as political leaders, department heads and project leads, as well as the teams who designed, developed, implemented and managed the AI solution, and the customer service employees who used the AI solution in their everyday work. Their trust was demonstrated in securing and sustaining the resources and support (e.g., funding, expertise, engagement) required for the AI system development, its implementation and ongoing maintenance and use.

External stakeholders included regulators, citizens affected by the system's decisions (as represented by key representatives such as the Ombudsman), and government departments that relied on Revenue NSW for debt collection. As trustors, these

TABLE 1 | Stakeholders whose trust was sought and required for the AI solution to be accepted.

Stakeholder group	Why engendering their trust was important
Political leaders and internal government employees who develop, operate, govern or have accountability for the AI solution	
Political leaders, department heads (sponsors) and project leads	As the leaders ultimately accountable for the AI-enabled service, these stakeholders needed to be reassured of the trustworthiness of the AI system and those charged with its development, implementation and management in order to make the initial and ongoing investment decisions and provide ongoing political support for the system.
Developers of the AI solution	Developers share some accountability for the trustworthiness of the AI solution, particularly its accuracy, reliability, and adherence to laws. They were motivated to innovate with a solution would be implemented in a benevolent way.
Users of the AI solution (i.e., domain experts and case workers)	As users of the AI solution in their everyday work of making case decisions and managing customer interactions for which they were accountable, these stakeholders needed to be assured the solution was fit for purpose and supported meaningful outcomes for customers. Their acceptance of the solution underpinned its sustained adoption. They also played a role in monitored its impact on customers (i.e., the public).
External stakeholders who regulate, rely on, or are subjected to the AI-enabled service	
Regulators (e.g., privacy and information commissioners)	Regulators act as conduits of justice in society, protecting citizens against wrongdoing and harm, including by public entities. They needed to be assured of the trustworthiness of the AI solution and the Department's implementation, management and governance of the solution in service provision, and its alignment with the public interest and privacy and data laws. In the absence of trust, they can publicly challenge and instigate inquiries.
Government departments reliant on Revenue NSW for debt collection (e.g., for taxes and fines)	These stakeholders needed to trust and accept the AI-enabled service as an effective alternative form of debt collection for their agency: '*...we have about 200 stakeholders that refer us debts- – councils, New South Wales Police, courts...and their expectation is we undertake enforcement action to collect those debts*' (P51).
Citizens affected by the AI solution's decisions and their representatives (e.g., citizen advocates, Ombudsmen)	The decisions informed by the AI-solution impact citizens, including vulnerable people. Citizen advocates and Ombudsmen interface with the public and act to safeguard their interests. Distrust of the AI solution and its implementation could erode public support and create pressure to stop the system, as occurred in the Robodebt case, and hamper broader government digitalization initiatives.

Note: A key part of the Ombudsman's role is to 'safeguard citizens from government actions which could adversely affect them' and 'give citizens a voice' to complain where they may not otherwise feel safe or comfortable to do so. They are often the only avenue readily available to individual citizens, particularly vulnerable citizens, seeking recourse on matters of maladministration or misconduct by public agencies (Neave 2015).

stakeholders needed to be sufficiently assured of the trustworthiness of the AI-enabled service and the Department's implementation of the AI solution to accept and approve its use.

There were several indicators of trust from these external stakeholders. An important behavioural indicator of citizen trust and acceptance of the solution is citizen complaints. One of the reasons for developing the solution was the rising number of complaints due to the garnisheeing process. These complaints 'declined significantly'—by 87%—once the AI solution was implemented, which the Ombudsman attributed to 'the application of the [machine learning] vulnerability model' (NSW Ombudsman 2024, 12).[4] Consequently, the Ombudsman formally discontinued his investigation, stating he was 'satisfied by the actions taken' by Revenue NSW (see NSW Ombudsman 2024, 15).

The agencies that relied on the Department for debt collection were also key stakeholders. Their sustained trust in the

Department's solution was evidenced behaviorally by their preparedness to share sensitive customer data with the Department to enable the development and continuous improvement of the AI solution, and their ongoing acceptance of the model as part of the debt collection process.

The solution not only survived the scrutiny of the Ombudsman but also the scrutiny of Parliamentary discussions and media attention in 2021 and continues to be used and recognised as an exemplar of trustworthy AI deployment (Digital NSW 2021).

5 | Approaches to Demonstrate Trustworthiness and Support Trust in Revenue NSW'S AI

A combination of approaches was taken by those responsible to navigate these challenges with the aim of demonstrating the trustworthiness of Revenue NSW's AI solution and its implementation to stakeholders to support their trust. In this section, we outline the key approaches (see Table 2), and for each, explain what the approach entailed, how the approach demonstrated trustworthiness and facilitated trust, as well as the considerations and recommendations on using these approaches in other contexts.

5.1 | Ensuring a Benevolent, Customer-Centric Purpose for AI Use

The foundations for building trust in any AI system begin by reflecting on whether AI is indeed suitable for addressing the problem at hand. This requires the organisation to have a deep understanding of the problem it is trying to solve, the affected stakeholders' situations, and how solving the problem will create benefit or value for stakeholders. With this understanding, the organisation can build trust by defining a benevolent purpose for the AI solution: a purpose that maximises positive outcomes for customers and affected stakeholders, while minimising potential harm. This approach builds trust by demonstrating *benevolence*—that is genuine care and concern for the welfare of customers and affected stakeholders.

This approach was taken at Revenue NSW by first gaining an evidence-based understanding of the problem. As the collector of revenue on behalf of over 250 agencies, an evidence-based understanding was important for Revenue NSW to justify and convince these agencies that an alternative approach to debt collection was needed. Its analysis of data uncovered several critical limitations of the existing debt collection processes. First, Revenue NSW found that relentlessly pursuing people who could never pay was both labour-intensive and ineffective. The existing processes relied on individuals (or their advocates) identifying themselves as vulnerable, which occurred after debt recovery had commenced when the impact of enforcement was high as well as the likelihood of needing to refund (NSW Government 2021c; NSW Ombudsman 2021a). In addition, manual examination of the records of over 3 million customers, 46 000 of whom are identified as vulnerable, was not financially viable (NSW Government 2021c). AI was seen as a way to counter these limitations because with appropriate data access, it 'can look at 900,000 cases every single day, it keeps covering ground much quicker' (P19).

TABLE 2 | Approaches to demonstrate the trustworthiness of the Department's AI solution.

Approach	Description
# 1: Ensuring a benevolent, customer-centric purpose for AI use	AI design, development and use are focused on solving customers problems and driven by a genuine desire to 'do good' enhance their welfare
# 2: Radical honesty with key stakeholders and seeking feedback	Open and transparent communication with stakeholders, providing honest information about the AI solution and actively seeking stakeholder feedback, including from customer representatives/advocates
# 3: Creating the solution with diverse input	Co-creating the solution with diverse input from multidisciplinary experts (e.g., domain, data, technical and legal)
# 4: Developing, validating and monitoring the AI solution	Rigorous and transparent development, validation and ongoing monitoring of the AI solution to ensure it operates as intended
# 5: Preserving human oversight and decision-making discretion	Ensuring human discretion is retained in consequential decisions that impact humans, with the ability to challenge AI decisions
# 6: Aligning the authorising environment	Ensuring AI use is aligned with and supported by the broader authorising and governance environment

Having a benevolent, customer-centric purpose and 'using artificial intelligence for good' (P70) was seen as critical to securing trust and gaining 'very, very broad support from our privacy group, from our data security team, risk and assurance and all those areas' (P51). This participant described that gaining trust 'was incredibly easy because we had the customer at the center of what we were trying to achieve here. Our goal was to reduce harm' but 'it would have been a completely different story if we wanted to build a model to target garnishee orders', that is engage in punitive behaviour.

The initiative for the AI solution was motivated by other actions the Department was taking which demonstrated they were taking a benevolent customer-centric approach. One leader claimed it was 'this whole big package of work...the AI is just one element of it...has enabled us to build the trust of some of the agencies' (P70). This 'big package' included streamlining write-off

processes and introducing a minimum protective amount in a bank account to 'remove the situation that you could be left with zero dollars in your account' (P51) after garnisheeing debts. They also introduced non-financial alternatives to resolving debt, such as participating in community support programs, including education, drug rehabilitation, and psychological programs, with the intention of achieving better long-term individual and community outcomes.

Revenue NSW also invested in understanding the problem from the perspective of vulnerable people, which led to more informed development of the AI solution, increasing the likelihood that the solution would be effective and trusted: 'naturally you start building smarter solutions to it, which is data driven' (P85). An example was looking into cases where the Department had written off a number of debts for young people. They interviewed people under 18 years old and talked to employees about their experiences. They used that research base to communicate with other stakeholders and then formulate a problem statement based on what they had learned, engaging with this cohort of people to explore alternatives.

Having a benevolent purpose was also important for demonstrating trustworthiness to Revenue NSW employees at all levels. Previously, officers had been dealing with fraught people, a 'confronting experience when you are at the other end of a call' (P34). The new solution enabled them to call people before debt escalated, and with alternative options. As the organisation shifted its mindset to how to help vulnerable people, momentum was built within the Department. The idea of using AI came from the data engineers, who had discovered that they already had the data to work out what was taking up to 2 years to realise with manual processes. In turn, a leader with responsibility for both the compliance (enforcing debt collection) and hardship (allowing refunds) teams understood their explanation of how this would benefit vulnerable people and provided the authority to continue development. The leader observed that:

> Lo and behold, other people got wind of it and then you should have seen the presentations and the awards... AI for good and not evil where we're actually removing the vulnerable customers out of an automated process.
>
> (P70)

Compliance and hardship teams began to work together to discuss and support vulnerable customers in an empathetic manner to find a program that would help them, given their individual needs.

Embedding a benevolent purpose into the service delivery through a fail-safe AI design further built trust by mitigating risks and removing the potential for harm to customers. The solution's design ensured that neither false positives nor false negatives would disadvantage customers in any way. Specifically, if the AI made a mistake in classifying someone as likely to be vulnerable, the worst that could happen was that the person's account would not be garnisheed for outstanding debts: 'The machine made the actions that were considered positive and never punitive and just stopped the punitive actions... So, the only real risk was that [...] we would stop sanctions for people who should get sanctions' (P19). This approach zeroed out any potential negative effects on citizens associated with AI's autonomous actions.

There are caveats to this approach. Taking a benevolent approach raises the question of benevolence towards whom? A government agency could claim benevolence by using an AI solution that genuinely helps some stakeholders (e.g., families) while disadvantaging others (e.g., singles). Hence, developing a trust-enhancing benevolent purpose requires a multi-stakeholder perspective, considering and balancing responsibilities to all affected parties, and ensuring that serving one stakeholder group is not at the expense or harm of another (Hurley et al. 2013).

5.2 | Radical Honesty With Key Stakeholders and Seeking Feedback

Even if the AI system's purpose is benevolent, ensuring widespread acceptance and legitimacy requires demonstrating *integrity* by upholding commonly expected values and norms in the delivery of the solution, including honesty, fairness and transparency. This was achieved by Revenue NSW by being honest and transparent about the AI solution, including its strengths, limitations, and performance, and consulting widely with affected stakeholders to ensure their views, concerns and input on the AI solution were considered early in the design phase and were effectively responded to. Doing this helped foster stakeholders' trust and a shared sense of ownership and acceptance of the solution.

Revenue NSW's leaders took the approach of 'radical honesty' when communicating to stakeholders about their AI solution, describing how they were upfront and transparent about the solution's capabilities and limitations during the development process: 'we're very, very transparent at every level and nobody has ever really pushed back' (P19). It was a significant shift from a common government culture that discourages bad news to being encouraged to 'talk about what happens when things go wrong' (P9).

Revenue NSW initially mapped all stakeholders who would be affected by or influential in the AI solution and planned how best to involve them throughout the process. These included regulators, privacy and information commissioners, other agencies, and, importantly, citizen representatives such as the Ombudsman—to ensure the voice of the people they were trying to help. They regularly met with the Ombudsman, Legal Aid[5] and union delegates and 'shared our dirty laundry' (P70). As P57 explained, 'we brought people together, and we talked about what we were doing and the outcomes that we had seen, and we had good, robust, open debate about the challenges'. The Department continually asked these stakeholders, 'What didn't we think of?' (P19) and encouraged them to voice concerns and problems early in the process so they could be addressed during the design and development stages to arrive at a solution that would be broadly trusted and accepted. It also demonstrated benevolence and integrity by transparently conducting robust impact assessment processes to identify potential risks, and putting

in place control mechanisms to mitigate these risks. Such effective consultation and responsiveness to stakeholders' concerns were acknowledged, for example, by the NSW Ombudsman: 'Revenue NSW worked responsively with us over time to ensure that its garnishee system operated more fairly, by taking account of vulnerability and situations of hardship.'

While the solution was being developed, there were cross-government initiatives to establish an AI assurance framework and governance forum, which included external experts in AI and human rights. As a demonstration of transparency and expression of radical honesty, Revenue NSW opened itself up to scrutiny by becoming one of the first trial cases to both test and be tested by the new assurance framework. This pushed them beyond what they called 'tick a box' assurance—'oh yeah, I have a committee. My committee doesn't understand anything, but I have a committee'—to be open to substantive interrogation by experts based on 'the substance of my ability to act properly in that capacity' (P19). One of the assurance framework committee members singled this Department out for building trust by taking accountability for the prior issues with fine garnisheeing, acknowledging 'Yep, mea culpa. This was us. This was the tool. We did it. We need to go back and fix it' (P34).

At the outset, there were mixed levels of trust among various groups of Department employees. Some 'were fully on board with machine learning' (P70) as they wanted automation to reduce the volume of repetitive administrative work in their roles. Others were 'very negative about automation' due to concerns about job loss (P70). The leader of one of the teams concerned about job security realised the approach needed to change to bring employees on the journey:

> We need to have an honest conversation with our workforce, not "this is just going to add value, it's not going to take away your job" type stuff, because I think that creates distrust. The volumes certainly put within revenue are growing exponentially, so it actually probably isn't going to take away their jobs, but the nature of work will change over the next 10 to 15 years. And understanding how AI influences that and having an honest conversation is the one that we need to have... with the people that are affected.
>
> (P36)

Knowing their team was 'very negative about automation', a leader initiated fortnightly face-to-face sessions with employees called 'Just Inform Me, Just Ask Me' so 'JIMJAMS' (P70) to build early awareness and address trust challenges as they emerged, encouraging employees to raise their questions and concerns. When the 'tricky questions' were asked, honest responses were provided, such as 'I don't know the answer to that yet – we're going to build that together'. This was perceived as a cultural shift from 'things being imposed on them and being told something at the very end' (P70).

Similarly, P36 described that part of her role was to build trust through 'mitigating risk, making sure that we'd understood at each touchpoint what was happening ... understanding those social impacts, both on the immediate workforce and on society in general'. Her stance was that 'change resistance is actually a very positive thing because it shows where the fear is' and 'the trust issues are about how is this going to affect me?'. Tailoring the approach and communications to address these impacts for employees was integral to supporting their trust and acceptance of the solution.

While radical honesty aims to signal trustworthiness by demonstrating integrity, it is not an easy or straightforward approach to securing trust, particularly in the risk-averse context of the public sector. At Revenue NSW, trust levels were impeded when transparency revealed problems to stakeholders in the early stages of the project. However, over time, Revenue NSW's open and honest approach achieved stronger stakeholder buy-in and trust, and the willingness to take accountability for the AI solution:

> This [radical honesty] initially actually took the levels of trust back a little bit by design...but then that's the short tail. The long tail of that is that people actually trusted the honesty of the process and the honesty of what came out of this...so that hard piece ... meant that the longer piece, the longitudinal piece was easier...There was a reason to have trust.
>
> (P19)

5.3 | Creating the Solution With Diverse Expert Input

Good intentions and purpose can be undermined if they are not matched by the *ability* to develop an AI solution that is fit for purpose in effectively addressing the business problem. Being able to develop a good AI solution requires effective collaboration between technical competence, legal expertise, data insights, and domain knowledge.

Revenue NSW brought together three expert teams required to competently execute the AI project: the analytics team (with a deep understanding of the data having been embedded in the business of fines and debt collection for many years), the business team (deep domain expertise on customers and vulnerability), and the digital team (expertise in system design and development). Having experienced data scientists working alongside seasoned caseworkers who had been taking the calls from people claiming to be experiencing hardship was critical for effectively designing the solution:

> [the data scientists] would sit and they would say, "Well, how do you manually figure out if someone is experiencing domestic violence, is homeless, has recently come out of a corrections centre? How do you separate who is and who isn't genuinely financially vulnerable?" And they would listen to calls with people. They would look at the notes. They would basically be very present in that process.
>
> (P57)

Having diverse team input was important in engendering the trust of senior leaders, who had the ultimate accountability to the public and government. It gave them assurance, for example, when questions such as 'how do we know that it's performing to its intention?' were answered by the people who were engaging with customers, not just the data scientist who was tweaking the model (P85).

The leaders themselves intentionally involved a range of stakeholders in various forums, to ensure they had carefully thought through potential impacts and to engender the trust of those stakeholders. For example, a cross-government governance mechanism was established, with representatives from police, corrective services, youth justice, LegalAid, and not-for-profits and justice organisations. The intent was that the governance was 'not just a machine learning thing' but instead a whole-of-business initiative centered on having an ethical approach on 'how were we going to do something different for individuals flagged as vulnerable' (P57). The Department sought input from experts in policy, law, privacy, cyber-security, risk, ethics, and human rights. Bringing these various sources of expertise together gave the senior leaders confidence that the solution would be effective and would be capably and responsibly developed.

Strengthening the Department's ability to deliver on its benevolent promises involved an intentional choice to build the solution in-house. The Department assigned an accountable executive: the person responsible to the Department head for the AI solution. This role was important for keeping abreast of developments, being able to answer questions as they arose, and keeping the Department honest by continually questioning 'Are we still good?' (P19). This executive sponsor held regular meetings and briefings with stakeholders, including employees and government regulators, which were important for building their trust in the governance processes. Leveraging internal capability and expertise meant that the accountable executive was close to the development process, giving stakeholders confidence in the solution's efficacy and accuracy rather than 'pay someone to build you something and then we all sit here going, "well it's great but we actually don't know how it works"' (P51). In-house development also enabled experimentation by removing the time and resource constraints common to outsourced technology projects.

It should be noted that bringing together diverse data science teams and domain experts is not always straightforward, as they tend to speak 'different languages' (Waardenburg et al. 2022). This requires a significant commitment from the organisation and can require specific coordination mechanisms to make it work smoothly. While Revenue NSW's decision to create the solution in-house ensured they developed the necessary internal capability, this is not always feasible in the resource-constrained public sector. Rather, the decision between in-house and outsourced AI development requires careful contextual consideration of the merits and limitations of each option.

5.4 | Developing, Validating and Monitoring the AI Solution Rigorously and Transparently

Creating the solution is only the starting point. Achieving trustworthiness also requires rigorous testing, validation and ongoing monitoring and refinement, pre- and post-deployment to demonstrate the *ability* and *integrity* of the AI solution and ensure models are sufficiently accurate and reliable for their assigned purpose.

More specifically, at Revenue NSW, people needed to be convinced of the AI solution's *ability* to correctly identify something as subjective as vulnerability—and that the training data was sufficient. During model development, key characteristics of vulnerability were refined through an iterative process of consulting domain experts, drawing on research on vulnerability, and analysing where there were differences in judgement between the AI model, the rules-based system that had been in operation,[6] and the domain experts' assessment of vulnerability:

> We really had some interesting scenarios when we did some testing around the model saying they are vulnerable, but the system says they're not. Or the system says they're vulnerable, but the model is saying they're overwhelmingly not vulnerable. And getting feedback from business experts like – so what are the conditions do you think that caused those scenarios - that allowed us then to improve the model.
>
> (P51)

Another ability-enhancing approach of Revenue NSW was the establishment of ongoing monitoring, refinement and continuous improvement processes to ensure the AI solution continued to be fit for purpose and reliably produce accurate outcomes, as well as increased in accuracy over time. To prevent problems associated with model drift, they used a static, rigorously tested baseline model in service delivery. While the static model was running daily operations, another dynamic model was continuously learning in the background. They regularly compared the dynamic model to the baseline model in a 'champion-challenger scenario' (P51)—initially every few weeks and then at least every 3 months—and when the dynamic model achieved higher accuracy with an acceptable level of explainability, the Department would make it the new static baseline model:

> Every couple of weeks, we retrain a model and see if the world has drifted away from what the algorithm already knows. We don't necessarily put in the new model that's drifted because we keep a model static... but when we have introduced new data elements to make it more precise, we would then also compare how much of a drift it has created and then literally check it and inspect it and sample it and say, "Is this right?" And what that feeds into is an internal discussion and briefing notes.
>
> (P19)

This approach was important for retaining trust and human control. It harnessed the power of AI's continuous learning ability to continually improve the model without compromising integrity. Over time, this resulted in an accuracy rate of 96%[7] (NSW Ombudsman 2024).

Major efforts were put into reassuring stakeholders of the AI solution's *integrity*—from the design, through the development, to post-implementation: 'here's the issue, here's the assurance framework, here's the experts we brought, here's the way we assessed negative matches' (P32). In the design, the Department's developers intentionally chose not to use inscrutable neural networks, despite their tendency to be highly accurate and efficient. Instead, they opted for an explainable model with traceable decision trees, justifying this decision based on the need to build trust through transparency and explainability: 'if you're going to build trust, you need to be transparent in explaining how you got to the decision' (P32).

Furthermore, while 'the full fines data set is used for this solution to ensure full inclusion and diversity in the data', data that does not assist the model in predicting vulnerability was excluded to ensure that it did not influence the model's predictions (NSW Government 2021c, 5). For example, a conscious decision was made to exclude cases from training where there was insufficient data to determine vulnerability, for example:

> [a] customer, never seen them before, they've just shown up in our system two months ago. They got one fine, never contacted us, and they've gone and got a Work Development Order or something like that. We really don't know a lot about that person to really call them either vulnerable or not.
>
> (P51)

In addition, the Department was aware that with this solution, 'our decisions have a behavior bias or discriminate to assist [the] vulnerable' which led to the realisation that removal of bias would reduce the positive outcomes (NSW Government 2021c, 5). Conscious decisions were made to not include attributes like gender and ethnicity in the attributes being considered to determine vulnerability—even though they were present in the dataset—because they did not improve predictive accuracy, reduced the risk of future accidental inclusion in training, and reduced the risk of human bias in the review process (NSW Government 2021b). There was also awareness of overfitting, where the model performs well on the data it was trained on, but poorly on new and unseen data. To address this risk, the Department trained on two-thirds of the data, and used the other third to evaluate the model's accuracy, and over time, reduced the number of attributes being used to predict vulnerability to improve generalizability (NSW Government 2021b).

The central data agency for government encouraged Revenue NSW to think through questions such as 'Are we getting the right false positive and false negative rates? Do we understand the consequences of false positives, false negatives if we're doing prediction type activities? And have we appropriately balanced or created ways of reversing any potential harm that comes from a false positive or a false negative?' (P9). They found the false negative rate was high; that is, where the model inaccurately assesses someone as vulnerable, partly because 'we don't know everything about everyone real time' (NSW Government 2021b, 17). Conversely, it was very accurate at predicting someone who was not vulnerable (NSW Government 2021b). This informed how the model was implemented in workflow decisions. For example, proceeding to enforcement for customers predicted not to be vulnerable, and picking up false negatives in the manual review process. The Department continues to regularly rerun technical checks compared to the baseline and validate the results with humans, including true positives, true negatives, false negatives, and false positives (see Appendix B for further details on the training and development approach).

During deployment, the Department was constantly measuring and monitoring the effectiveness and achievement of the intended outcomes, and adapting accordingly. For example, 'how many we're doing, how many are successful, what is the revenue being generated, how many customer contacts, we look at complaints' (P51). Post-implementation, the Department produces monthly reports to continuously assess the efficiency of the service:

> How many people on a monthly basis had we identified as vulnerable, who was in the backlog, the hardship team – how many people were they able to reach out to, what's the number of hours that people on working development programs had been credited with, what does that mean in terms of dollar amount, are people sticking in with the [programs] or are they kind of falling off partway through.
>
> (P57)

It also has annual reviews with its external stakeholders, partly for education and partly to review how the process works. A learning from the early stages of the project was that these approaches were critical to building and maintaining trust with the business areas.

While Revenue NSW had a rigorous approach demonstrating trustworthiness and enhancing trust, it can be resource-intensive and challenging to execute in resource-poor organisations, which may counterintuitively stop the development of value-creating AI projects. Hence, leaders need to find a suitable balance between rigor and feasibility when developing AI, taking into consideration the context and risks involved.

5.5 | Preserving Human Discretion in Decision-Making

Machine-learning algorithms are only as good as the data used to train them, and they struggle with the contextual understanding of humans' individual situations (Sarker 2021). Even the most rigorously developed AI systems cannot always be trusted to make decisions on people's unique circumstances. Hence, in high-risk contexts, or contexts where AI is being used to inform consequential decisions, retaining human discretion over decision-making is a powerful trust-enhancing approach that is often necessary to reassure stakeholders of the trustworthiness (i.e., *ability*, *integrity*, and *benevolence*) of decision-making processes. As P51 explained: 'We still want a person to make that assessment because we have to recognise the model's not going

to get it right all the time.' In some contexts and jurisdictions, such human oversight is required to comply with legal or ethical frameworks.

At Revenue NSW, the collective learning from years of legal scrutiny of the automated garnishee system was: (1) a fully automated system was illegal, (2) an automated system with human oversight but no human decision-making was legal but morally 'wrong', and (3) automation could legally be used to support and augment human decision-making on garnisheeing, for example, by identifying the accounts eligible to be garnisheed before the decision, and to request banks to garnishee accounts after the decision. The bottom line was that human oversight that did not involve a mental reasoning process to form a judgement was insufficient. The AI part of the solution, to detect and remove vulnerable people from automated garnisheeing, needed to involve human discretion:

> Legal advice was sought, and it was determined that [the solution] cannot be fully automated. That has to have a human intervention loop...because we're not ready nor do we have the legislative framework to go further for a straight-through machine process.
>
> (P57)

Hence, an augmented approach was used to determine vulnerability, which ensured accountability remained with case officers:

> This solution is designed to make predictions that assist Revenue NSW Officers to make decisions. The AI does not make any decisions itself. Accountability for the decisions to direct customers to alternative resolution pathways remains entirely with officers within Revenue NSW.
>
> (Digital NSW 2021)

Specifically, the AI removes any customers it predicts as vulnerable from the list of accounts to garnishee, but ultimate decisions are subject to human judgement and discretion: 'They left human beings in a loop and said, "If you don't agree with AI, override it, you are human, you make the decision."' (P36). As P19 reinforced: 'A human being has the last laugh. The machine never overrides a human, a human overrides the machine, and it normally settles people's nerves'.

This augmented human decision-making approach preserved relationships and trust between customers and service providers, which is key to meaningful work and employee engagement. The augmented design not only ensured the solution's legal integrity and decision-making ability but also reinforced the benevolent customer-centric approach by mitigating the risk of harm to customers: if the AI solution made a mistake, there was still a layer of human review to ensure accuracy and a contextualised human understanding of the customer:

> The machine made the actions that were considered positive and never punitive and just stopped the punitive actions without asking a human being. But then after the fact, or even every day when people could review what is being stopped, they were allowed to intervene either way.
>
> (P19)

Accountability for decisions, with the assistance of the AI solution for vulnerable customers, remained with the officers (Revenue NSW 2021a).

Supporting this process was an investment the government had made in improving the decision-making frameworks to ensure the right level of flexibility, accountability and fairness. This involved educating employees on discretionary judgements, an important factor in 'human-in-the-loop' decisions. The aim was to have consistent and fair decision-making for circumstances the legislation does not cover; for example, someone overstaying paid parking at a hospital due to a procedure taking longer than planned. Good judgement and decision-making would say, 'Well, we're going to issue a caution rather than a fine in that instance' (P57).

While retaining humans in the loop is trust-enhancing, we note that the importance and practicability of retaining human decision-making discretion depend on the use case, regulations, and context. In some cases, close human involvement would encumber the process to the extent that AI's efficiency benefits are lost. Such situations require careful assessment of relevant legislation and the trade-offs between the cost of human involvement and the potential risk of AI errors. In cases posing minimal risk and potential for harm, AI may be trusted without human oversight or discretion (for instance, recommender engines).

5.6 | Aligning With the Authorising Environment

Finally, ensuring the *integrity* of AI solutions in public services requires empowerment and support from the *structural assurances* provided by the broader authorising environment, aligned with internal governance. This includes formal authorization from budget holders, delegated decision-makers, and legislation, as well as formal and informal authority granted through ministers, commissioners, central agencies, and other influential stakeholders (APSC 2021). It also includes *jurisdictional clarity*, that is, clear lines of responsibility and an identifiable locus of authority to enable citizens to hold the correct decision-makers to account (Hobolt et al. 2013).

Being aligned with regulation and laws is a necessary foundation for AI solutions to be perceived as trustworthy and legitimate: '...for trust, we need the legislation to allow us to utilize AI' (P36). While aligning with legislation and regulation has implications for the design and implementation of AI systems (as discussed in #4 and #5), it may sometimes be necessary to proactively advocate for changes in the authorising and governance environment to meet stakeholders' needs and expectations of public sector AI deployment (e.g., decision-making delegation). The rigor involved in changing legislation can reassure stakeholders that appropriate laws, regulation, and external scrutiny are in place to protect people and ensure AI solutions are safe and trustworthy.

Revenue NSW proactively ensured, and had independently audited, the alignment of the solution, policy, and legislation as a critical trust-building measure. The accountable owner for the solution described how their own trust came from this assurance, and the integration between the AI solution, the policy and the other delivery work related to delivering outcomes, rather than being technology-led, or using AI as a point solution. The Department also sought legal advice on the solution and its use of AI. This step was particularly important in the long shadow cast by Robodebt and for the concerns the Ombudsman had with automated garnisheeing. It enabled the Department to counter distrust by politicians and the media and for stakeholders who were not 'digitally savvy' (P57).

In addition, Revenue NSW independently requested changes to the legislation to remove its own enforcement powers related to the collection of government revenue and fines. These changes involved a combination of measures designed to engender trust, such as removing the power Revenue NSW had to imprison fine defaulters and requiring a qualified social worker to oversee development programs. One leader explained that legislating clear, simple rules was important to clarify 'we won't do this, and we won't do that,' thereby combating any claims of inappropriate use (P32). Executive leaders expressed the importance of any AI-enabled program aligning with the legal framework and the organisation's risk appetite:

> My strong view is technology is an enabler, it's not the lead. So as we work through with the expert panels what legislation change might look like, then we can put together a package of work that will help us leapfrog to solutions.
>
> (P36)

In addition to aligning with the broader authorising environment, Revenue NSW strengthened its internal governance. This included articulating and publicly committing to its objectives, planned outcomes and principles, including that 'data does not leave Revenue NSW boundary', that 'data and actions are restricted to authorized officers and processes', that 'consent is legislative' and that 'the existing solution would not be used for another purpose without rethink and revalidation' (NSW Government 2021c, 6–8). The Department formalised the senior leaders who had accountability for the process, established monthly meetings, with reporting from the analytics and Fines team on outcomes, issues and suggestions for future improvements. As well as the measures outlined in #3, the people ultimately accountable to the minister and public cited this forum as significant for engendering their trust. Specialists from both the Fines and Debts, and Analytics teams provided oversight of the process, so that Revenue NSW retained accountability and decision-making rights, not the AI (NSW Government 2021c).

Throughout the development, Revenue NSW instituted new governance and auditing functions to mitigate the risks associated with AI. This included a steering committee with representatives of impacted stakeholders, an outcome-based ministerial approval process, and conducting regular independent third-party audits. The intent was to ensure there was an auditable and defensible record of decision-making, as well as to have appropriate 'checks and balances in place...there is some safety net' (P72).

Aligning the authorising environment can be challenging given the evolving nature of laws and regulations, particularly across departments with different legislative requirements. Moreover, pursuing changes to laws and regulations is desirable when those changes benefit citizens and society, as was the case with changes requested by Revenue NSW. However, governments might also use this approach to legitimise the harmful use of an AI system (for instance, to violate citizens' privacy or discriminate against specific populations). Hence, efforts to change existing regulations in connection with AI systems should be subjected to broad consultation and diligent ethical scrutiny. We can also learn from this case that clarity of roles and responsibilities within and across agencies enhances coordination and information sharing, supporting trust and participation among public-sector collaborators (Gil-Garcia et al. 2019).

6 | Reflections on Trustworthy AI in the Public Sector

Given the low-trust environment in which this AI solution was incubated, the outcomes from this case are commendable. Assessments of trustworthiness do not exist in a vacuum. Rather, in the context of AI, trustworthiness should be understood in relation to the rapid proliferation of digital technologies across public and private sectors and the vulnerabilities these technologies create for people (Lockey et al. 2021). AI technology is not neutral but reflects the values and worldviews of its designers (Martin 2019). Its effectiveness is typically enabled by large-scale collection of personal data, which can be invasive and obscure to the subjects of data collection, raising questions about surveillance (Zuboff 2019), data justice (Masiero and Das 2019), and digital inclusion (Stelmaszak et al. 2024). The *datafication* of people combined with algorithmic decision-making enabled by AI technologies risks negative consequences (Newell and Marabelli 2015; Rinta-Kahila et al. 2022) and can potentially amplify existing power imbalances (Zuboff 2019). In this context, the Revenue NSW case portrays how these challenges can be addressed to enable the solution to be fully deployed and used by central government regulators as an exemplar of trustworthy AI. The case reinforces that a key source of 'good reasons' to trust (Lewis and Weigert 1985; Mollering 2006, 13) comes from structural assurances and control mechanisms, including laws, governance, and oversight mechanisms.

To deepen insight on the nature and effectiveness of Revenue NSW's approaches to engendering trust, in the next section, we systematically contrast these approaches with those taken by another government agency, in the high-profile Robodebt program which resulted in widespread distrust.

6.1 | Comparison With Robodebt

The Robodebt system was originally implemented in response to a political party's campaign promises to balance the budget and improve the integrity of the welfare system. Reducing the national debt had been a long-standing objective and campaign

theme of Australian political parties, and one means to this end has been to enforce welfare compliance. This strategy had often been accompanied by the *demonization* of welfare support recipients (Hutchens 2022), with politicians claiming that some citizens intentionally *defraud* the welfare system by collecting more payments than they are entitled to. While the agency in question had a long history of managing complex IT projects, the Robodebt system was developed and deployed hastily without heeding best practices. The system replaced a previous, human-centric process with poorly designed automation that shifted the burden of proof by forcing the targeted citizen to prove they did not owe a debt to the government.

Four similarities between the Revenue NSW and Robodebt cases make it a relevant comparison. Both solutions involved an Australian government agency leveraging algorithmic decision-making to collect revenue from individual citizens. Both agencies had evidence of significant outstanding debt to the government. Both systems became operational around the same time, with the Robodebt system implemented in 2016 and Revenue NSW AI 2 years later. Both had the lawfulness of their schemes challenged by individuals, advocacy groups, regulators, and the media.

However, our comparative case analysis revealed key differences in the approaches taken in relation to (a) the solution's purpose, (b) stakeholder orientation, (c) system design, and (d) relationship with the authorising environment, which led to radically different trust outcomes. A summary of these different approaches is shown in Table 3.

6.2 | Considering Unintended Consequences

The case comparisons (above) highlight how the trust-supporting approaches we identified can facilitate trustworthiness. However, it is prudent to consider that under different conditions some of these approaches may have unintended consequences. For example, the approaches assume the government's perspective on human rights and justice aligns with the broad interests of society. When this is absent, such as with the Robodebt system, it can enable government leaders to frame and defend flawed programs as being in the public interest, even when they demonstrably cause harm and result in mistreatment of some citizens. This highlights the importance of balancing the needs of individual citizens with the needs of the collective in the delivery of public services to avoid alienating segments of the population and undermining trust in public services (Commonwealth of Australia 2024; Halma and Guetzkow 2023).

Second, radical honesty may backfire if weaknesses or potential problems are leaked to the media early in the consultation, design and development stages, before they have been addressed. Careful and thoughtful stakeholder management and consultation are required across all phases of the AI lifecycle to prevent such leaks and potential damage to trust. Third, while having a human in the loop can protect people from harm by identifying and correcting AI mistakes, this approach may create a false sense of security if the human actors are not effectively able to intervene in the process at relevant stages, or if they grow excessively reliant on AI and fail to exercise their own judgement (Rinta-Kahila et al. 2023). For instance, the leaders defending the Robodebt system claimed that the system's decisions were subject to human discretion; however, this was not meaningful, as the staff were instructed to accept the algorithmic calculations and discourage citizens from contesting them. Fourth, while Revenue NSW pursued legislative changes to protect the public from potential AI harm, a less trustworthy organization might abuse its authority and advocate for legal changes that are not in the public's interest to legalise destructive uses of AI.

6.3 | Building Foundational Capabilities for Trustworthy AI Solutions

Across our larger dataset, we observed a pattern in which the agencies that deployed AI solutions perceived as trustworthy had each first invested in developing foundational capabilities in data and customer-centricity over several years. For Revenue NSW, these investments enabled moving from benevolent customer-centricity as a broad organisational ethos to individual-level action. The investment established a function for strategy and transformation, building capability in customer experience, strategic planning and project governance, and where the 'biggest thing was the transformation around how the organization interacted with the citizen' (P57). The change involved training programs, awards that incentivised customer-centric behaviour, reshaping processes and systems to present a unified front to customers, ensuring communication with customers was empathetic, clear and in plain English and gaining a deeper understanding of customers through the investment in data. As one leader described, in the years preceding the AI program's implementation, the Department was 'living and breathing data every day' (P70). A restructure to increase customer-centricity then brought previously disparate teams together, which, combined with the data analytics capability, provided fertile ground for the development of the AI solution.

These foundations supported trust in the AI solution in two key ways. First, the deep understanding employees had of the Department's customers and service processes, along with the measures they had in place, enabled the AI solution to be developed in a way that catered to the complex and diverse range of individual customer circumstances. It built employee capability for problem-solving in a way that would not have been possible in procuring an AI solution from the market:

> With the sort of deep problem solving that requires a really, really detailed understanding of the business processes...we felt we could do a better job than the [consulting firms].
>
> (P57)

The data also provided the Department with the means to effectively govern the AI solution. An executive explained the level of data they had access to was a critical success factor, as they 'knew at all times how many fines had been issued, how many fines had been paid, how many fines had gone through reminder notice, how many enforcement order, how many people are paying an enforcement order versus not' (P57).

TABLE 3 | A comparison between Robodebt and Revenue NSW's approaches.

Dimension	Approach	
	Robodebt system	**Revenue NSW AI solution**
Purpose	*Economic purpose*: Designed to maximise revenue to the government despite negative human implications *Punitive*: Sent letters to people, including vulnerable citizens, claiming they owed the government money *High-risk and negative impact on citizens*: The system had a substantial impact on citizens, requiring them to either pay or challenge the alleged debt; false positives impacted citizens negatively and unfairly, subjecting them to distress	*Human purpose*: Designed to find ways to protect and help people, and counter external-to-agency concerns of losing revenue *Protective*: Prevented the garnisheeing of accounts for people likely to be vulnerable *Low-risk and positive impact on citizens*: The system could only have a positive effect on citizens. Identifying a person as vulnerable meant that the citizen would be protected from automatic garnisheeing; false positives had no negative impact beyond potential perceived unfairness of some being exempt from garnisheeing while others were not.
Stakeholder orientation	*Closed*: Failed to consult key stakeholders, such as Digital Transformation Agency and Australian Council of Social Service, in the development of the solution; ignored criticism when the system was in use	*Open*: Engaged openly with different stakeholders, striving to identify blind spots in the Department's approach; exposed the solution to scrutiny
System design	*Simplistic*: Rule-based algorithmic system that made deterministic and unrealistic assumptions about overpayments to citizens based on inadequate data *Human out of the loop*: The algorithmic system made decisions that directly affected citizens, with no corrective human intervention *No feedback mechanism*: No oversight mechanisms established; the system was kept operational for 3 years despite ample evidence of inaccuracy	*Elaborate*: Machine-learning system that produced dynamic, probability-based predictions about citizens' vulnerability status *Human discretion*: The AI solution made decisions that directly affected citizens, but those were reviewed and could be overruled by human employees at any point *Strong feedback mechanism*: Extensive training, validation, and regular ongoing monitoring of outcomes
Relationship with the authorising environment	*Contradictory*: The solution operated in contradiction to legislation. Internal and external stakeholders' concerns about the solution's legality were ignored or dismissed; the deployment agency increased punitive powers without legislative support	*Aligned*: The solution operated in alignment with legislation. Revenue NSW sought legal advice and consulted heavily with regulators to ensure its approach was legal; the Department successfully requested changes to the legislation that curbed their punitive powers

Second, the investment in data management and governance mechanisms enhanced the willingness of other agencies to share data with Revenue NSW, providing a greater pool of high-quality citizen data for developing and training the AI model and ensuring its accuracy. These data governance mechanisms included appointing data custodians, conducting impact assessments, understanding the relevant regulations, establishing data sharing agreements, and minimising unnecessary collection of personal information from citizens.

7 | Management Implications

The question we asked at the outset of this research was: how can government leaders demonstrate to stakeholders their trustworthy use of AI in public services? Our case analysis showcases *how* and *why* a public sector agency—grappling with significant trust challenges—managed to design, develop and deploy a widely accepted AI solution through the use of six trust-supporting approaches. The considerations and caveats with each approach demonstrate that trust challenges, and the approaches to build trust, extend beyond the technology artefact itself and into broader organisational practices and regulatory frameworks. Trust in an AI system cannot be properly evaluated or achieved unless the institution implementing AI commands stakeholder confidence. Revenue NSW nurtured trustworthiness at multiple levels—technical, organisational, and regulatory—underscoring the intricate tapestry of public sector accountability structures. A phased rollout, systematic engagement with different stakeholders, and openness to external scrutiny contributed symbiotically to demonstrating trustworthiness and supporting trust.

This rendering of trust as multidimensional, socio-technical and socio-political in nature, rather than a static metric, is a key insight that reminds us that demonstrating trustworthiness is fluid, contextual and involves iterative learning. As

such, trust challenges cannot be solved in isolation, and there can be no formula for building trust. However, we outline three important implications for public sector leaders seeking to demonstrate trustworthiness and support trust in their agency's use of AI.

7.1 | Demonstrating Trustworthy AI Use Requires Relational and Technical Capabilities

An important learning from this case is that demonstrating trustworthiness is as much relational as technical (Gillespie & Dirks, 2026). Certainly, leaders need to understand and manage the risks and vulnerabilities raised by unique technical attributes of AI, including its autonomous capabilities, dynamic learning, and inscrutable operating logic. However, addressing these technical trust challenges alone is not sufficient, particularly in the socio-political context of government. A proactive, intentional strategy for building and sustaining stakeholders' trust by demonstrating the trustworthiness of the AI solution and its integration into service delivery is needed, as well as demonstrating to stakeholders the broader trustworthiness of the organisation deploying and governing the AI solution (Söllner et al. 2016; Van der Werff et al. 2021). If stakeholders are cynical and sceptical of the deploying organisation, even the most trustworthy AI application may be rejected.

Supporting relational trust requires leaders to think carefully about the stakeholder ecosystem: who needs to be involved, who can influence, and who will be impacted by the AI solution. This stakeholder ecosystem often includes regulators, funding agencies, ministers, other public agencies, citizens, employees, and people with a vested interest in the existing or improved service. Taking stock of trust in the relationship with each stakeholder group, understanding their needs, vulnerabilities and concerns in relation to the AI solution, demonstrating genuine care and concern for their interests, and investing time in genuinely addressing their concerns are examples of important relational approaches to supporting trust with stakeholders. This investment in relationships and commitment to understanding stakeholders can be mutually reinforcing, as genuine consultation typically leads to improved solutions that are more trustworthy, which reinforces trust and acceptance.

7.2 | Trustworthy AI Implementations Are Problem-Driven, Not Technology-Driven

Another important insight from this case and the other AI cases we studied is that trustworthy AI implementations are problem-driven, not technology-driven. As one participant reflected: 'One of the reasons that this was an award-winning policy and I think the key to its success...is it wasn't an "AI project". It was a project that AI was part of.' (P36) We observed trust and adoption were more difficult to secure in AI projects driven chiefly by the technical capabilities of AI (e.g., the productivity benefit of fully automating a decision-making process) or solely financially motivated (e.g., a desire for cost savings or revenue generation), as opposed to being driven by a genuine problem affecting stakeholders (e.g., using AI to identify and remove human bias in a service). We frequently heard of projects failing due to leaders being 'hoodwinked' into AI-led service transformations based on unrealistic and untested promises of productivity gains. In contrast, a real problem coupled with a benevolent purpose is poised to build trust by signalling that the Department is deploying the AI for the right reasons.

An important implication is that leaders need to be mindful of the purpose of their AI initiatives and what problems they are seeking to solve with AI. One approach is to start with a 'hairy chestnut'—something that has been a significant problem for customers and that the Department has been unable to solve through traditional approaches. Such problems mobilise and motivate diverse stakeholders to work together to innovate and overcome concerns, and critically, evaluate whether AI is the 'right tool' for solving the problem. It is important to acknowledge that AI solutions are rarely neutral, but rather embody the values and priorities of their designers (Martin 2019). Having a meaningful problem to solve, combined with a benevolent purpose and a customer-centric approach, helps ensure that AI solutions embody and reflect trust-inducing values.

7.3 | Demonstrating Trustworthy AI Use Invariably Involves Navigating Tensions

An implication of our case study, evident in the considerations and recommendations, is that demonstrating trustworthy development and use of AI solutions is not straightforward and typically involves navigating several tensions. Addressing one trust challenge can exacerbate other trust issues. For example, more comprehensive datasets can augment the accuracy of AI solutions, with accuracy a key determinant of the trustworthiness of the system. However, this must be balanced with consent to collect and use this data, and individual rights to privacy. Knowing too much about an individual—for example, their medical or criminal records or the number of properties they own—may compromise the individual's privacy and trust in the government. A related tension is that more accurate AI solutions based on deep learning models are often less explainable than less accurate but more explainable models. Another trust tension evident in the case relates to openness and transparency. Leaders being open with stakeholders about the limitations of the AI solution can have short-term negative impacts on trust and can seem counterintuitive in the risk-averse context of government. However, in the long term, this can be more effective in managing expectations and building trust.

A tension in government is that the business case to justify AI solutions can reveal inefficiencies, bias and inaccuracies in the existing manual processes, opening the potential for political and reputational ramifications. Revenue NSW described how it 'shared its dirty laundry' with existing stakeholders, a stance that required courage and humility. This was significant in 'building the trust of all the agencies that Revenue New South Wales [deal with]', balancing its remit to collect debt with 'trying to bring issuing authorities along that journey with us as well' (P70).

While some of these tensions may exist with traditional technology solutions, they are exacerbated with AI solutions because of AI's unique attributes and its capability to make or inform decisions affecting stakeholders. The implication for government leaders is that implementing trustworthy AI is not a linear, set-and-forget journey. It requires judgement calls to navigate tensions in a way that supports trust, and ongoing efforts to ensure that trustworthiness is demonstrated to sustain trust over time.

8 | Boundary Conditions and Future Research

The strengths of our case study approach need to be considered in light of its limitations. While our case documentation and media articles are time-stamped and gave insight into the unfolding of events, our interview data was retrospective. While the interviews benefitted from participants' reflection on longer-term impacts of approaches to supporting trust, they may suffer from limitations of memory bias and retrospective sense-making. Like all case studies, there is a trade-off between deep and rich insights, and the ability to generalise findings (Yin 2018).

A second limitation is the lack of direct measurement of public trust in the AI-enabled service. As noted in Appendix A, for privacy reasons, we could not identify citizens who were subjected to the AI solution. Rather, we reference objective data, quantifying the significant reduction in customer complaints since the introduction of the AI solution. We also inferred public trust and acceptance from media articles, interviews and reports by stakeholders whose role involved advocating for customers and representing the public interest. Future research could employ surveys and focus group methodologies using representative sampling to examine citizen trust of AI-enabled services directly.

An important and promising line of future research is to examine how the application of normative frameworks like 'trust by design' in the development and implementation of AI solutions may reduce power asymmetries, embed accountability checks, and minimise harms in each step of the AI lifecycle. While Revenue NSW's AI solution met the AI ethics framework subsequently deployed by the Government, we call for further research that directly examines whether and under what conditions 'trust by design' principles and responsible AI assurance frameworks promote stakeholder trust.

The case was positioned in the Australian government context, at a time of significant distrust in the government's use of automated technology in public service delivery. Translation and generalizability of the trust-building strategies to broader contexts, including other countries, the private and non-profit sector, and other AI use cases, warrant examination. We expect that with careful adaptation and customization to fit the local political, social and cultural context, the six trust-building approaches will be broadly applicable across a range of contexts and applications. This is supported by our broader research which found these approaches applied in multiple other public sector AI use cases. It also aligns with prior work suggesting the underlying theoretical mechanisms and principles supporting trust, such as demonstrating ability, benevolence and integrity, and assuring trust through institutional safeguards, are applicable across a variety of cultural contexts, levels, and referents of trust (Ferrin and Gillespie 2010; Fulmer et al. 2024; Gillespie and Dirks 2026).

The generalizability of the trust supporting approaches across cultural contexts is further reinforced by recent multi-country surveys of public trust and attitudes towards AI. These works reveal strong convergence and endorsement in public expectations for AI governance and regulation across countries (Gillespie et al. 2023, 2025). For example, across 48,000 people representing 47 countries, 83% report a greater willingness to trust AI systems when governance mechanisms are in place that align with the six trust supporting approaches, such as regular system monitoring for accuracy (approach #4), human intervention to correct, override or challenge AI recommendations (approach #5), and laws, regulations and policies to govern responsible AI use (approach #6; Gillespie et al. 2025). This consensus was found despite significant variation across countries in the level of public trust and acceptance of AI use, and regulatory frameworks. Furthermore, a 17-country study reported that 95% of people endorse the principles and practices of trustworthy AI—which underpin the six trust supporting approaches—as important for their trust in AI systems (Gillespie et al. 2023). These findings held across a range of AI use cases (e.g., AI use in healthcare, human resources, security, and recommender systems), suggesting their universal relevance (Gillespie et al. 2023; 2025).

We anticipate that in contexts characterised by high mistrust, limited legal and institutional safeguards, and high-stakes AI applications, stakeholders will require greater reassurance that AI solutions are being designed, developed and deployed in a trustworthy and responsible way, and may demand additional trust-building approaches beyond those identified in the current study. Similarly, AI use cases that heighten particular vulnerabilities, such as the security of sensitive information, are likely to require specific practices to secure trust (e.g., privacy by design, reassurances of data security).

Recent evidence suggests that people in advanced economies (as defined by the International Monetary Fund), such as the United States, are generally less trusting and accepting of AI use, and more likely to believe that existing regulation and institutional safeguards are inadequate to make AI use safe, compared to countries with emerging economies, such as India (Gillespie et al. 2023, 2025). An implication is that trust-supporting approaches—such as aligning the authorising environment—may be more important for securing trust in advanced than in emerging economies. The implementation of the comprehensive European Union AI Act[8] will continue to reshape the authorising environment and institutional safeguards governing AI, clarifying expectations of what organisations developing and deploying AI are required to do to demonstrate trustworthiness through adherence to regulatory frameworks.

Advances in AI technologies may also reveal new challenges for trust that require adaptations to existing trust-building practices. Given the rapid advances and uptake of generative AI technologies and large language models, we call for future research that examines the efficacy of the recommended strategies in the context of high-stakes generative AI use.

Acknowledgments

We would like to thank Revenue NSW and our research participants for their time and insight. This research was supported by a University of Queensland Research Support grant and the University of Melbourne-KPMG Chair in Trust grant awarded to the second author. Tapani Rinta-Kahila's research receives support from the Research Council of Finland (370017) and the Australian Research Council (DE240100269).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Endnotes

[1] Interviews are an accepted method to examine how trustworthiness is demonstrated and how trust is built and sustained over time; see Lyon et al. (2015). *Handbook of Research Methods on Trust*. Edward Elgar Publishing.

[2] The opportunity cost was estimated to be $200 000 AUD.

[3] Ironically, Revenue NSW's AI solution—which was instigated to address these concerns by preventing garnisheeing accounts of vulnerable people—initially got caught in the line of fire: it was reportedly misunderstood by some as amplifying the practice of automatic garnisheeing Revenue NSW. (2021b). *Using machine technology: Clarification on claims in media in response to the ombudsman report 'The new machinery of government: using machine technology in administrative decision-making'*. Revenue NSW, https://www.revenue.nsw.gov.au/news-media-releases/using-machine-technology.

[4] Actionable garnishee complaints reduced from 214 in the 2015–2016 year prior to the introduction of the AI solution to 28 in 2019–2020 post implementation of the AI solution NSW Ombudsman. (2024). *Revenue NSW—The lawfulness of its garnishee order process*. NSW Government https://www.ombo.nsw.gov.au/Find-a-publication/publications/reports-to-parliament/other-special-reports/the-new-machinery-of-government-using-machine-technology-in-administrative-decision-making.

[5] LegalAid is a government funded organisation which provides free or low-cost legal services to people who cannot afford a lawyer.

[6] The "system" is used by employees for debt management, and has functionality including looking up customer debt, sending messages to debtors, setting up payment plans and receiving payments.

[7] "...when tested against 250 000 Revenue NSW customers in 2018, the model was found to be 96% accurate in terms of (automatically) identifying people as vulnerable that Revenue NSW itself would otherwise (manually) have assessed as vulnerable" NSW Ombudsman. (2024). *Revenue NSW—The lawfulness of its garnishee order process*. NSW Government https://www.ombo.nsw.gov.au/Find-a-publication/publications/reports-to-parliament/other-special-reports/the-new-machinery-of-government-using-machine-technology-in-administrative-decision-making.

[8] The EU AI Act is one of the most significant reforms to legislation and regulation of AI which governs members of the European Union.

References

Andrews, P., T. de Sousa, B. Haefele, et al. 2022. "A Trust Framework for Government Use of Artificial Intelligence and Automated Decision Making." *Cornell University arXiv*, 2208.10087, 1–20.

APSC (Australian Public Service Commission). 2021. "Understand Your Authorising Environment." https://www.apsc.gov.au/initiatives-and-programs/workforce-information/taskforce-toolkit/governance/understand-your-authorising-environment.

Asatiani, A., P. Malo, P. Nagbøl, E. Penttinen, T. Rinta-Kahila, and A. Salovaara. 2020. "Challenges of Explaining the Behavior of Black-Box AI Systems." *MIS Quarterly Executive* 19, no. 4: 259–278.

Asatiani, A., P. Malo, P. Nagbøl, E. Penttinen, T. Rinta-Kahila, and A. Salovaara. 2021. "Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems." *Journal of the Association for Information Systems* 22, no. 2: 325–352.

Attorney-General's Department. 2011. "Australian Administrative Law Policy Guide." Australian Government. https://www.ag.gov.au/sites/default/files/2020-03/Australian-administrative-law-policy-guide.pdf.

Bedué, P., and A. Fritzsche. 2022. "Can We Trust AI? An Empirical Investigation of Trust Requirements and Guide to Successful AI Adoption." *Journal of Enterprise Information Management* 35, no. 2: 530–549.

Bennett, P. 2018. "Opening Government: Shaping Democratic Outcomes in the Information Age." In *Opening Government: Transparency and Engagement in the Information Age*, edited by J. Wanna and S. Vincent. ANU Press.

Berente, N., B. Gu, J. Recker, and R. Santhanam. 2021. "Managing Artificial Intelligence." *MIS Quarterly* 45, no. 3: 1433–1450.

Birch, S., and N. Allen. 2010. "How Honest Do Politicians Need to be?" *Political Quarterly* 81, no. 1: 49–56.

Colquitt, J. A., and J. Rodell. 2011. "Justice, Trust, and Trustworthiness: A Longitudinal Analysis Integrating Three Theoretical Perspectives." *Academy of Management Journal* 54, no. 6: 1183–1206.

Commonwealth of Australia. 2024. "COVID-19 Response Inquiry Report." Australian Government. https://www.pmc.gov.au/sites/default/files/resource/download/covid-19-response-inquiry-report.pdf.

Corbin, J., and A. Strauss. 2015. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 4th ed. SAGE Publications.

DCS. 2022. "Department of Customer Service Annual Report 2021–2022." NSW Government. https://www.nsw.gov.au/sites/default/files/2022-12/DCS-annual-report-2021-2022_0.pdf.

Department of Finance. 2024. "National Framework for the Assurance of Artificial Intelligence in Government: A Joint Approach to Safe and Responsible AI by the Australian, State and Territory Governments." Australian Government. https://www.finance.gov.au/government/public-data/data-and-digital-ministers-meeting/national-framework-assurance-artificial-intelligence-government.

Desouza, K., and G. Dawson. 2023. "Pathways to Trusted Progress With Artificial Intelligence." IBM Center for the Business of Government. https://www.businessofgovernment.org/report/pathways-trusted-progress-artificial-intelligence.

Devine, D., V. Valgardsson, W. Jennings, G. Stoker, and H. Bunting. 2025. "The Causes of Perceived Government Trustworthiness." *European Journal of Political Research* 64, no. 3: 1394–1412.

Digital NSW. 2021. "Using Artificial Intelligence to Identify and Support Customers Facing Hardship." NSW Government. https://www.digital.nsw.gov.au/article/using-artificial-intelligence-to-identify-and-support-customers-facing-hardship.

DTA. 2024. "Policy for the Responsible Use of AI in Government." Australian Government. https://www.digital.gov.au/policy/ai/policy.

Dubois, A., and L. Gadde. 2002. "Systematic Combining: An Abductive Approach to Case Research." *Journal of Business Research* 55, no. 7: 553–560.

Dwivedi, Y., L. Hughes, E. Ismagilova, et al. 2021. "Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy." *International Journal of Information Management* 57, no. 2021: 101994.

Eisenhardt, K. 1989. "Building Theories From Case Study Research [Case Study]." *Academy of Management Review* 14, no. 4: 532–550.

European Commission. 2022. "European AI Excellence and Trust in the World: Event Report." Dubai 2022. European Commission. https://excellenceandtrust.intouchai.eu/wp-content/uploads/2022/05/GFA-European-AI-Excellence-and-Trust-in-the-World_16052022-1.pdf.

Faraj, S., S. Pachidi, and K. Sayegh. 2018. "Working and Organizing in the Age of the Learning Algorithm." *Information and Organization* 28, no. 1: 62–70.

Ferrin, D., and N. Gillespie. 2010. "Trust Differences Across National-Societal Cultures: Much to Do, or Much Ado About Nothing?" In *Organizational Trust: A Cultural Perspective*, edited by M. Saunders, D. Skinner, G. Dietz, N. Gillespie, and R. Lewicki. Cambridge University Press.

Fulmer, A., D. Ferrin, E. Denison, and N. Gillespie. 2024. "Culture and Trust in Organizational Contexts." In *The Oxford Handbook of Cross-Cultural Organizational Behavior*, edited by M. Gelfand and M. Erez. Oxford University Press.

GAO. 2021. "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities." US Government Accountability Office. https://www.gao.gov/products/gao-21-519sp.

Gil-Garcia, J., A. Guler, T. Pardo, and G. Burke. 2019. "Characterizing the Importance of Clarity of Roles and Responsibilities in Government Inter-Organizational Collaboration and Information Sharing Initiatives." *Government Information Quarterly* 36, no. 4: 101393.

Gillespie, N., and N. Daly. 2025. "Repairing Trust in Public Sector Agencies." In *Handbook on Trust in Public Governance*, edited by F. Six, J. Hamm, D. Latusek, E. Van Zimmeren, and K. Verhoest. Edward Elgar Publishing.

Gillespie, N., and K. Dirks. 2026. "Trust Dynamics in Organizations." In *The Oxford Handbook of Organisational Social Evaluations*, edited by R. Younger and A. Zavyalova. Oxford University Press.

Gillespie, N., S. Lockey, and C. Curtis. 2021. "Trust in Artificial Intelligence: A Five Country Study." The University of Queensland and KPMG Australia. https://espace.library.uq.edu.au/view/UQ:e34bfa3.

Gillespie, N., S. Lockey, C. Curtis, J. Pool, and A. Akbari. 2023. "Trust in Artificial Intelligence: A Global Study." The University of Queensland and KPMG Australia. https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf.

Gillespie, N., S. Lockey, T. Ward, A. Macdade, and G. Hassed. 2025. "Trust Attitudes and Use of Artificial Intelligence: A Global Study 2025." The University of Melbourne and KPMG. https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html.

Gioia, D., and E. Pitre. 1990. "Multiparadigm Perspectives on Theory Building." *Academy of Management Review* 15, no. 4: 584–602.

Grimes, M. 2006. "Organizing Consent: The Role of Procedural Fairness in Political Trust and Compliance." *European Journal of Political Research* 45, no. 2: 285–315.

Halma, M., and J. Guetzkow. 2023. "Public Health Needs the Public Trust: A Pandemic Retrospective." *BioMed* 3, no. 2: 256–271.

Hamm, J., G. Möllering, and K. Darcy. 2024. "Integrating Focal Vulnerability Into Trust Research." *Journal of Trust Research* 14: 237–255.

Henley, J. 2021. "Dutch Government Resigns Over Child Benefits Scandal." https://www.theguardian.com/world/2021/jan/15/dutch-government-resigns-over-child-benefits-scandal.

Hobolt, S., J. Tilley, and S. Banducci. 2013. "Clarity of Responsibility: How Government Cohesion Conditions Performance Voting." *European Journal of Political Research* 52, no. 2: 164–187.

Holmes, C. 2023. "Royal Commission Into the Robodebt Scheme." Commonwealth of Australia. https://robodebt.royalcommission.gov.au/system/files/2023-09/rrc-accessible-full-report.PDF.

Hurley, R., N. Gillespie, D. Ferrin, and G. Dietz. 2013. "Designing Trustworthy Organizations." *MIT Sloan Management Review* 54, no. 4: 74–82.

Hutchens, H. 2022. *The Humanization of 20th Century Europe's Perpetrators: How Humanizing Our History's Perpetrators Can Better Our Future*. East Carolina University.

Kalesnikaite, V., and K. Baker. 2025. "Private Organization's Moral Behaviour and Citizen Support for Public-Private Partnerships: Evidence From a Survey Experiment." *Public Management Review* 27, no. 6: 1563–1587.

Kelly, S., S. Kaye, and O. Oviedo-Trespalacios. 2023. "What Factors Contribute to the Acceptance of Artificial Intelligence? A Systematic Review." *Telematics and Informatics* 77: 101925.

Laux, J., S. Wachter, and B. Mittelstadt. 2024. "Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk." *Regulation & Governance* 18, no. 1: 3–32.

Lewis, J., and A. Weigert. 1985. "Trust as a Social Reality." *Social Forces* 63, no. 4: 967–985.

Lindebaum, D., V. Glaser, C. Moser, and M. Ashraf. 2023. "When Algorithms Rule, Values Can Wither." *MIT Sloan Management Review* 64, no. 2: 1–5.

Lockey, S., N. Gillespie, D. Holm, and I. Someh. 2021. "A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions." 54th Hawaii International Conference on System Sciences, Hawaii. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/a08c7344-3c5b-4b1b-8782-5ba791dad6d6/content

Long, C., and S. Sitkin. 2018. "Control–Trust Dynamics in Organizations: Identifying Shared Perspectives and Charting Conceptual Fault Lines." *Academy of Management Annals* 12, no. 2: 725–751.

Lundin, M. 2007. "Explaining Cooperation: How Resource Interdependence, Goal Congruence, and Trust Affect Joint Actions in Policy Implementation." *Journal of Public Administration Research and Theory* 17, no. 4: 651–672.

Lyon, F., G. Möllering, and M. Saunders. 2015. *Handbook of Research Methods on Trust*. Edward Elgar Publishing.

Mansoor, M. 2021. "Citizens' Trust in Government as a Function of Good Governance and Government Agency's Provision of Quality Information on Social Media During COVID-19." *Government Information Quarterly* 38, no. 4: 101597.

Manzoni, M., R. Medaglia, L. Tangi, C. Van Noordt, L. Vaccari, and D. Gattwinkel. 2022. *AI Watch Road to the Adoption of Artificial Intelligence by the Public Sector: A Handbook for Policymakers, Public Administrations and Relevant Stakeholders*. European Commission. https://publications.jrc.ec.europa.eu/repository/handle/JRC129100.

Margetts, H., and C. Dorobantu. 2019. "Rethink Government With AI." *Nature* 568, no. 7751: 163–165.

Martin, K. 2019. "Designing Ethical Algorithms." *MIS Quarterly Executive June* 18, no. 2: 129–142.

Masiero, S., and S. Das. 2019. "Datafying Anti-Poverty Programmes: Implications for Data Justice." *Information, Communication & Society* 22, no. 7: 916–933.

Mayer, R., J. Davis, and F. Schoorman. 1995. "An Integrative Model of Organizational Trust." *Academy of Management Review* 20, no. 3: 709–734.

McKnight, D., M. Carter, J. Thatcher, and P. Clay. 2011. "Trust in a Specific Technology: An Investigation of Its Components and

Measures." *ACM Transactions on Management Information Systems* 2, no. 2: 1–25.

Medaglia, R., J. Gil-Garcia, and T. Pardo. 2023. "Artificial Intelligence in Government: Taking Stock and Moving Forward." *Social Science Computer Review* 41, no. 1: 123–140.

Merriam, S., and R. Grenier. 2019. *Qualitative Research in Practice: Examples for Discussion and Analysis*. John Wiley & Sons.

Misra, S., S. Sharma, S. Gupta, and S. Das. 2023. "A Framework to Overcome Challenges to the Adoption of Artificial Intelligence in Indian Government Organizations." *Technological Forecasting and Social Change* 194: 122721.

Mollering, G. 2006. *Trust: Reason, Routine, Reflexivity*. Emerald Group Publishing.

Neave, C. 2015. "Exploring the Role of the Commonwealth Ombudsman in Relation to Parliament." Parliament of Australia. https://www.aph.gov.au/-/media/05_About_Parliament/52_Sen/Publications_and_resources/Papers_and_research/Papers_on_Parliament/pop63/c03.pdf.

Newell, S., and M. Marabelli. 2015. "Strategic Opportunities (And Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification'." *Journal of Strategic Information Systems* 24, no. 1: 3–14.

NSW Government. 2021a. "Case Study: Revenue NSW Uses AI to Help Vulnerable Customers." Revenue NSW. https://www.haveyoursay.nsw.gov.au/artificial-intelligence/case-study-revenue-nsw-uses-ai-to-help-vulnerable-customers.

NSW Government. 2021b. "Using Machine Learning & Data to Support Vulnerable Customers." AI Summit, NSW.

NSW Government. 2021c. "Vulnerable Customer Support Using Machine Learning." AI Summit, NSW.

NSW Ombudsman. 2021a. "Appendices. Revenue NSW - The Lawfulness of its Garnishee Order Process." NSW Government. https://cmsassets.ombo.nsw.gov.au/assets/Reports/Section-31-Special-Report-Revenue-NSW-Appendices.pdf.

NSW Ombudsman. 2021b. "The New Machinery of Government: Using Machine Technology in Administrative Decision-Making." NSW Government. https://www.ombo.nsw.gov.au/Find-a-publication/publications/reports-to-parliament/other-special-reports/the-new-machinery-of-government-using-machine-technology-in-administrative-decision-making.

NSW Ombudsman. 2024. "Revenue NSW – The Lawfulness of Its Garnishee Order Process." NSW Government. https://www.ombo.nsw.gov.au/Find-a-publication/publications/reports-to-parliament/other-special-reports/the-new-machinery-of-government-using-machine-technology-in-administrative-decision-making.

OECD.AI. 2023. "Policies, Data and Analysis for Trustworthy AI." Organization for Economic Cooperation and Development. https://oecd.ai/en/.

Office for Artificial Intelligence. 2022. "A Guide to Using Artificial Intelligence in the Public Sector." UK Government Digital Service. https://assets.publishing.service.gov.uk/media/6093f6bfe90e0726f7b69caf/A_guide_to_using_AI_in_the_public_sector__Mobile_version__V2.pdf.

O'Neill, O. 2018. "Linking Trust to Trustworthiness." *International Journal of Philosophical Studies* 26, no. 2: 293–300.

Ouattara, E., E. Steenvoorden, and T. van der Meer. 2023. "Political Trust as an Evaluation Against Normative Benchmarks? A Two-Wave Survey Experiment on the Role of Normative Benchmarks in the Evaluative Model of Political Trust." *International Journal of Public Opinion Research* 35: 1–13.

Park, S. 2020. "Multifaceted Trust in Tourism Service Robots." *Annals of Tourism Research* 81: 102888.

Patton, M. 1999. "Enhancing the Quality and Credibility of Qualitative Analysis." *Health Services Research* 34, no. 5 Pt 2: 1189–1208.

Pettigrew, A. 1990. "Longitudinal Field Research on Change: Theory and Practice." *Organization Science* 1, no. 3: 267–292.

Pew Research Centre. 2024. "Public Trust in Government: 1958–2024." Numbers, Facts and Trends Shaping Your World. https://www.pewresearch.org/politics/2024/06/24/public-trust-in-government-1958-2024/.

Prats, M., E. Phillips, and S. Smid. 2023. "Insights From the 2021 OECD Trust Survey: How People Evaluate the Trustworthiness of Government Institutions & Implications for Policymakers." *Behavioral Science & Policy* 9, no. 2: 9–20.

Revenue NSW. 2020. "Work and Development Order Annual Report 2019/20." NSW Government. https://www.nsw.gov.au/sites/default/files/2022-11/work-and-development-order-annual-report-2020-2021.pdf.

Revenue NSW. 2021a. "Using Artificial Intelligence to Identify and Support Customers Facing Hardship." https://www.digital.nsw.gov.au/article/using-artificial-intelligence-identify-and-support-customers-facing-hardship.

Revenue NSW. 2021b. "Using Machine Technology: Clarification on Claims in Media in Response to the Ombudsman Report." The New Machinery of Government: Using Machine Technology in Administrative Decision-Making. Revenue NSW. https://www.revenue.nsw.gov.au/news-media-releases/using-machine-technology.

Ring, P., and J. Perry. 1985. "Strategic Management in Public and Private Organizations: Implications of Distinctive Contexts and Constraints." *Academy of Management Review* 10, no. 2: 276–286.

Rinta-Kahila, T., E. Penttinen, A. Salovaara, W. Soliman, and J. Ruissalo. 2023. "The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation." *Journal of the Association for Information Systems* 24, no. 5: 1378–1412.

Rinta-Kahila, T., I. Someh, N. Gillespie, M. Indulska, and S. Gregor. 2022. "Algorithmic Decision-Making and System Destructiveness: A Case of Automatic Debt Recovery." *European Journal of Information Systems* 31: 1–26.

Rothstein, B., and J. Teorell. 2008. "What Is Quality of Government? A Theory of Impartial Government Institutions." *Governance* 21, no. 2: 165–190.

Rousseau, D., S. Sitkin, R. Burt, and C. Camerer. 1998. "Not So Different After All: A Cross-Discipline View of Trust." *Academy of Management Review* 23, no. 3: 393–404.

Russell, S., and P. Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Pearson.

Sarker, I. 2021. "Machine Learning: Algorithms, Real-World Applications and Research Directions." *SN Computer Science* 2, no. 160: 1–21.

Shao, Z., Y. Guo, X. Li, and S. Barnes. 2020. "Sources of Influences on Customers' Trust in Ride-Sharing: Why Use Experience Matters?" *Industrial Management & Data Systems* 120, no. 8: 1459–1482.

Shollo, A., K. Hopf, T. Thiess, and O. Müller. 2022. "Shifting ML Value Creation Mechanisms: A Process Model of ML Value Creation." *Journal of Strategic Information Systems* 31, no. 3: 101734.

Söllner, M., A. Hoffmann, and J. Leimeister. 2016. "Why Different Trust Relationships Matter for Information Systems Users." *European Journal of Information Systems* 25, no. 3: 274–287.

Someh, I., B. Wixom, C. Beath, and A. Zutavern. 2022. "Building an Artificial Intelligence Explanation Capability." *MIS Quarterly Executive* 21, no. 2: 143–163.

Stelmaszak, M., E. Wagner, and N. DuPont. 2024. "Recognition in Personal Data: Data Warping, Recognition Concessions, and Social Justics." *MIS Quarterly* 48, no. 4: 1611–1636.

Tyler, T. R. 1998. "Trust and Democratic Governance." In *Trust and Governance*, edited by V. Braithwaite and M. Levi. Russell Sage Foundation.

Tyler, T. R. 2006. "Psychological Perspectives on Legitimacy and Legitimation." *Annual Review of Psychology* 57, no. 1: 375–400.

Van Bekkum, M., and F. Borgesius. 2021. "Digital Welfare Fraud Detection and the Dutch SyRI Judgment." *European Journal of Social Security* 23, no. 4: 323–340.

Van der Werff, L., K. Blomqvist, and S. Koskinen. 2021. "Trust Cues in Artificial Intelligence: A Multilevel Case Study in a Service Organization." In *Understanding Trust in Organizations: A Multilevel Perspective*, edited by N. Gillespie, C. Fulmer and R. Lewicki. Routledge.

van Erkel, P., and T. van Der Meer. 2016. "Macroeconomic Performance, Political Trust and the Great Recession: A Multilevel Analysis of the Effects of Within-Country Fluctuations in Macroeconomic Performance on Political Trust in 15 EU Countries, 1999–2011." *European Journal of Political Research* 55, no. 1: 177–197.

van Noordt, C., R. Medaglia, and L. Tangi. 2023. "Policy Initiatives for Artificial Intelligence-Enabled Government: An Analysis of National Strategies in Europe." *Public Policy and Administration* 40, no. 2: 215–253.

Waardenburg, L., M. Huysman, and A. Sergeeva. 2022. "In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms." *Organization Science* 33, no. 1: 59–82.

Weber, M. 1978. *Economy and Society: An Outline of Interpretative Sociology*. University of California Press.

Yin, R. 2018. *Case Study Research and Applications: Design and Methods*. 6th ed. SAGE Publications.

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

Appendix A

The Research Method

This study aimed to advance understanding of the trust challenges experienced in developing and integrating AI into public services, and the approaches that helped address these challenges by demonstrating trustworthiness and facilitating trust. We chose an exploratory, abductive, qualitative case study method, which is appropriate for research questions examining complex, context-dependent phenomena for which there is limited existing knowledge (Eisenhardt 1989; Yin 2018).

Our engagement with the Department commenced with the first interview in October 2021 and completed with a validation (member check) meeting in September 2025. After conducting a first meeting and interview with the Project Director, we mapped the various stakeholders of the AI-solution and those with in-depth understanding of the AI design, development, implementation and governance process based on documentary analysis and information provided by the Department. This stakeholder mapping was continually updated as understanding of the case developed and included the AI solution builders and developers, domain experts, accountable leaders, people in governance roles, employees who used the AI solution, and representatives of the public impacted by the solution. The Department facilitated access to stakeholders, which we supplemented with an independent snowballing technique, asking at the end of each interview who else we should interview to obtain diverse perspectives on the case.

Our interview data consists of 14 in-depth interviews, ranging in length from 27 to 78 min, with an average length of 57 min. This provided approximately 200 pages of transcripts. Our interviewees (see Table A1) represented a broad range of roles, including key people involved in the initiation, design, development, implementation, management, regulation, governance or use of the AI solution and project, including politicians, government regulators, senior leaders, managers, developers, data engineers and subject matter experts (SMEs) with case work experience. Interviews were conducted either in-person or online, were recorded and professionally transcribed, and continued until data saturation was achieved with no new themes emerging.

In addition to interviews, we collected rich case documentation, including media articles and press releases, agency presentations, strategy documents, annual reports, parliamentary and ombudsman reports, as well as communications on agency websites and social media (see Table A2). This documentation provided longitudinal, time-stamped data on the unfolding of the case, and importantly, information on external stakeholders' trust and broader views, concerns, praise and acceptance of the AI-enabled solution, including by representatives of the public interest, such as the Ombudsmen, Privacy and Information Commissioners. Table A3 outlines the strategies used to enhance the trustworthiness and rigour of the research method and analysis (Merriam and Grenier 2019).

For privacy reasons, we could not directly access citizens who were subject to the AI solution's decisions. Rather, citizens' trust and acceptance of the solution is inferred by the 'significant reduction' in citizen complaints about the Department's debt collection process once the AI-solution was implemented, as confirmed in interviews (P19) and by the Ombudsman's independent report (NSW Ombudsman 2024). A key role of the Ombudsman in Australia is to 'safeguard citizens from government actions which could adversely affect them' and 'give citizens a voice' to complain where they may not otherwise feel safe or comfortable to do so. They are often the only avenue readily available to individual citizens, particularly vulnerable citizens, seeking recourse on matters of maladministration or misconduct by public agencies (Neave 2015, 44). Hence, the reduction of citizen complaints as reported by the Ombudsman is an important indicator.

This behavioural indicator was triangulated with positive accounts of public sentiment and outcomes from interviews with domain experts who interacted with the public and regulators charged with upholding the public interest, triangulated by media reports and the Ombudsman's reports. For example, central government leaders cited

this case as an exemplar of trustworthy AI (NSW Government, 2021), and the Department received awards for the program as evidenced in public domain annual reports (Revenue NSW 2020). Designers of the NSW Government AI Assurance Framework described this use case as a success (P9).

In the first stage of analysis, the transcripts were read by four of the authors, and discussed through multiple meetings to identify key themes and patterns (Pettigrew 1990). The second stage was more structured and involved using NVIVO qualitative software and a coding protocol based on the emergent themes and patterns (Gioia and Pitre 1990) to code and triangulate the emerging insights from interviews with the secondary data sources. Using a process of systematic combining, we iterated between these data sources to make comparisons, refine terms and validate findings (Corbin and Strauss 2015; Dubois and Gadde 2002; Eisenhardt 1989). We synthesised interview transcripts and archival data into a case history and analysis of 42 401 words (42 pages), including a chronology and timeline of the AI solution development and implementation, summary of the perceived risks and benefits, the trust challenges experienced, and the approaches used to engender trust, each supported by quotes from the informants and case documents.

The analysis process resulted in identifying key trust challenges and approaches taken by the Department to address these challenges by demonstrating trustworthiness and supporting trust. We then conducted cross-case analysis with seven other cases within our broader research program, for further triangulation, identifying corresponding challenges and trust-supporting approaches. This further solidified our confidence with the findings reported in the Revenue NSW case.

TABLE A1 | Interviewee profiles and interview timing.

Role on project	#	Title	Interview timing (min)	Date of interview
Developer/builder/operator[a]	P32	Revenue NSW Senior Leader – Digital	27	July 2022
	P51	Revenue NSW – Data engineer and system developer	71	May 2022
	P57	Revenue NSW – Leader – Digital	52	April 2022
	P19	Revenue NSW Senior Leader – Data	78	October 2021
Domain expert	P70	Revenue NSW Leader – Business[b]	48	May 2022
Sponsors	P36	Revenue NSW Leader – Business lead and project sponsor	54	June 2022
	P85	Revenue NSW Senior Leader – Accountable executive (executive sponsor)	66	October 2022
Governance	P9	Central Government Leader – Data and AI Governance	54	March 2022
	P34	Central Government AI Subject Matter Expert (SME) – Governance forum	56	April 2022
	P58	Central Government AI SME – Governance forum	62	May 2022
	P76	Central Government Regulator – Privacy[b]	58	August 2022
	P77	Central Government Legal SME – Governance forum	52	September 2022
	P84	Central Government Regulator – Information[b]	63	September 2022
Governance/strategy	P72	Central Government Politician – Accountable owner	53	August 2022

[a]Involved in the design, development and operation of the AI solution.
[b]Informed understanding of public trust and sentiment, and the views of affected stakeholders.

TABLE A2 | Data sources used for case study analysis.

	Data source	How used in the case analysis
D1	Central government website describing – How is AI used to enhance customer interactions – Building a digital and customer capable workforce – Digital and customer capability framework – AI Assurance framework	Context on AI in government and governance mechanisms in place
D2	Revenue NSW website describing – the service (2 documents) – the Dept and its mission – the Dept's use of AI	Public communication on the service and Department's use of AI and data Inform case context Information on use of AI and data
D3	Revenue NSW strategy	Inform case context
D4	Revenue NSW annual report (x2)	Inform case context
D5	Ombudsman report and annexure[a]	Informed understanding of public trust, sentiment and concerns of public interest with use of AI in the Department's operations
D6	Media reports on Revenue NSW use of AI and media interviews with Revenue NSW (x6)[a]	Informed understanding of public trust and sentiment on the case
D7	Parliamentary transcripts[a]	Inform understanding of central government context, including funding and regulation of the AI solution and public interest considerations
D8	Opening Government: Shaping democratic outcomes in the information age (Bennett 2018)	Referred by P19. Approach to transparency and engagement in government
D9	Presentation to central government AI committee	Inform the case context, design approach and solution
D10	Revenue NSW press release addressing media reports	Inform community concern with Revenue NSW use of AI
D11	Program Annual report	Inform the context and content of overall solution
D12	Revenue NSW Website: Changes to legislation including the Fines Act	Removing powers to seek imprisonment of fine defaulters, to enable a social worker to supervise development program activities negotiated in lieu of paying fines, and the authority to get information from other agencies and financial institutions
D13	Social media update	Announcement of legislation changes on LinkedIn
D14	Government legislation	Legislation pertaining to the collection of government revenue and fines

[a]Informed understanding of public trust and sentiment and views of affected stakeholders.

Appendix B

Training and Development of the AI-Enabled Solution

In developing the model, Revenue NSW tested multiple different types of machine learning models and implemented what they found to be most accurate fit for their purpose, including 'regression, decision tree, gradient boost' and different languages such as 'R, Python and WEKA' (NSW Government 2021b, 17). Over time, Revenue NSW found that training/testing/validation on an entire large population, and the characteristics of that large population, are quite stable. Training/testing/validation against multiple random sample subsets has shown the same result for all subsets, giving confidence that data sampling uniformity exists and is a feasible basis for ongoing training and validation (P19).

In 2018, the model was found to be 96% accurate at automatically identifying people that staff would manually have assessed as vulnerable. A threshold was determined from analysis of collection rates relative to the prediction of vulnerability, finding a significant drop in the ability for debts to be repaid when there was over 70% likelihood that the person was vulnerable (NSW Government 2021b). The business rules for the garnisheeing system were then modified so that fine defaulters over a predicted vulnerability threshold were excluded from files to be garnisheed (NSW Ombudsman 2024).

To address concerns on the need for human oversight of the remaining accounts to be garnisheed, a Check Summary Report was introduced in 2019 using a traffic light system. If all traffic lights were green, the delegated officer approved the garnishee order. Vulnerability was one of several criteria used to exclude individuals from the garnishee list. Over a period of 4 years, changes were made to the Check Summary Report to address concerns that human oversight was merely 'rubber-stamping' and did not engage human discretion and judgement (NSW Ombudsman 2024, 25).

TABLE A3 | Strategies for enhancing the trustworthiness and rigour of the research method.

Strategy	Description
Triangulation	To gain a comprehensive understanding of the trust building strategies, we used method, data and investigator triangulation (Patton 1999). We collected data using interviews conducted with a range of stakeholders internal and external to the Department, triangulated with archival data from a diverse range of sources including Department presentations, reports and project documentation, and websites, as well as public domain reports, parliamentary transcripts, media articles and social media posts. The research team included six researchers that brought interdisciplinary expertise in the fields of management, information systems, psychology, ethics and public policy.
Member checks	Between May and September 2025, multiple members of the Department reviewed this paper and verified the accuracy and comprehensiveness of the case study description, the strategies, practices and actions taken, and their rationale. The case study was approved for publication by the Senior Executive responsible for the Department. Interviewees were provided with the opportunity to review and correct their interview transcript prior to analysis.
Analysis	Three members of the research team analysed the data, working iteratively until key themes were identified and agreed upon, and a coding protocol established and applied. Multiple meetings and discussions were used to clarify and refine the themes and codes, including gaining feedback from the broader research team on the coding protocol and trust-supporting approaches. Inter-coder reliability was further supported by having two members of the analysis team code the same sub-set of interviews and documents using the coding protocol and discuss any divergent codes and interpretation until consistent agreement was reached.
Researcher reflexivity	After each interview, the interviewer(s) wrote a memo reflecting on the conduct and content of the interview and their reflections and observations. The research team met regularly to reflect on the emerging findings and analysis.
Adequate engagement and variation in data collection	Across the research program, we interviewed 95 stakeholders across a range of government agencies. For the Revenue NSW case, we interviewed 14 internal and external stakeholders (see Table A1), complemented by a diverse range of documents (see Table A2).
Rich, thick descriptions	For each case, we wrote a thorough case description. The Revenue NSW case description, used as the basis for this case study paper, was 42 401 words (42 pages).