

School of Law
Centre for Policy Futures



CREATE CHANGE

Artificial Intelligence and Australian Charities

Governance, Bias, Human Rights and Public Trust for
Non-Profits in the Algorithmic Age

Kim D Weinert
Brydon T Wang



WHITE PAPER | MARCH 2026



Contents

Forewords 2
Executive Summary 4
Authors 6

AI, Charities and the New Trust Challenge 9

What Needs to Change 10

Why AI Has Arrived in the Australian Charity Sector 11

The Promise and the Pressure of AI in Charitable Work 12
What Charities Are Actually Doing 13

The Governance Blind Spot 16

Bias, Power and the Charitable Context 19

Human Rights, Public Benefit and the Fragility of Trust 21

Trustworthy AI Framework for Charities 24

'Trust' and 'Trustworthiness': Why the Difference Matters for AI in Charities 26
Benevolence: AI Must Demonstrably Serve Beneficiaries, Not Institutions 27
Integrity: Alignment with Charitable Purpose and Existing Laws 29
Ability: Charities Must Understand and Govern What They Deploy 30
Trustworthiness as a Governance Threshold 31

Principles for AI Governance in Charities 33

Principle 1: Charitable Purpose Must Retain Its Primacy 34
Principle 2: End-User and Beneficiary Voices Must Be Centred 35
Principle 3: Algorithmic Transparency and Explainability 36
Principle 4: Proactive Bias Identification and Mitigation 37
Principle 5: Meaningful Human Oversight and Override 38
Principle 6: Clear Accountability Allocation 39

Implementation 40

Conclusion 45

DOI 10.14264/5cdb683

Copyright © 2026 Kim D Weinert and Brydon T Wang. This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International <<https://creativecommons.org/licenses/by-nc-sa/4.0/>>.



The authors are grateful to Professor John Swinson for his thoughtful comments on an earlier draft, and to Professor Rain Liivoja for his guidance and support in the development and refinement of this White Paper. Any errors or omissions remain the authors' own.

We acknowledge the Traditional Owners and their custodianship of the lands on which our university stands. We pay our respects to their Ancestors and descendants, who continue cultural and spiritual connections to Country. We recognise their valuable contributions to Australian and global society.

Forewords



What Must Remain Human

The Honourable Michael Kirby AC CMG

Former Justice of the High Court of Australia

For more than four decades, I have observed and written about the interaction between law and emerging technologies. From the early incorporation of computers into legal administration to contemporary debates about artificial intelligence, a recurring question has persisted. How can legal systems adapt to the beneficial capacity of technological change without surrendering their fundamental commitments to justice, accountability and human dignity?

Artificial intelligence now presents that question in an especially acute form. No longer confined to internal optimisation or administrative assistance, AI increasingly mediates decisions that affect access to and monitoring of services, support and remedies. The emergence of more autonomous and agentic systems, capable of not just analysing information but of initiating actions, generating communications and shaping institutional responses, intensifies these concerns. When such systems influence decisions affecting vulnerable individuals, efficiency (whilst clearly relevant and useful) cannot be the sole measure of legitimacy.

Charities occupy a distinctive position within this landscape. They are institutions of civil society entrusted with public benefit functions, regulatory privileges and moral authority. They operate at the boundary between law, policy and human need. When charities adopt AI systems to allocate resources, prioritise assistance or assess eligibility, they exercise power that must be justified, constrained and capable of explanation. The consequences of error or exclusion in these settings are not abstract but become the lived experience of the vulnerable. Human beings have empathy, a feeling of association, beneficial motivation, love for others and kindness. Such characteristics, at the heart of the charitable instinct, are not (as such) present in the technology of machines.

Much of the current discourse on AI governance has focused on governments and private corporations. This White Paper makes a valuable and timely contribution by turning attention to civil society, where formal safeguards are often thinner and reliance by end-users and beneficiaries is often greater. It recognises that traditional legal duties, shaped around human decision-makers, may struggle to address systems that operate at scale, rely on historical data, lack empathetic motivation and obscure the reasons for particular outcomes.

The significance of this White Paper lies in its reminder that technological sophistication does not relieve institutions of responsibility. For charities, whose legitimacy rests upon public trust and the faithful pursuit of public benefit, the adoption of artificial intelligence must remain subject to principled scrutiny. However advanced such systems become, decisions affecting those in vulnerable circumstances must remain capable of justification and human accountability. In civil society, no less than in the courts, the exercise of power must remain attentive to dignity, equality and trustworthiness as well as human-like empathy, love and kindness.

This White Paper makes a thoughtful and timely contribution to that enduring challenge.

Why AI Governance in the Charitable Sector is a Human Rights Issue

Dr Lorraine Finlay

Human Rights Commissioner
Australian Human Rights Commission

We live in a society that is simultaneously demanding stronger human rights protections and also relying more heavily than ever on data-driven technologies to make life more efficient. It is tempting to assume that these ambitions inevitably conflict, particularly as artificial intelligence augments or replaces elements of human decision-making. But the real task is not to choose between innovation and rights. It is to ensure that the design, development and deployment of technology are structured in ways that uphold human dignity, fairness and accountability from the outset.

Human rights are not optional safeguards applied after innovation has occurred. They are the framework that governs how power is exercised. AI is increasingly influencing how charities determine access to services and support. When systems shape who receives assistance or how applications are prioritised, they are exercising power. That power must remain transparent, accountable and capable of human rights scrutiny.

Public debate about AI often centres on productivity and efficiency. In the charitable sector, however, we must also ask how these systems affect equality, fairness and access. Charities serve individuals facing structural disadvantage and who have limited alternatives and capacity to challenge decisions. When AI systems shape access to services in these contexts, they exercise real and consequential power over people's lives. The fact that this power is exercised by charities rather than government does not lessen its human rights significance.

Technology does not neutralise power. It can conceal it. Systems that appear objective may embed assumptions that disadvantage those already facing systemic barriers. Protecting human rights requires more than formal neutrality – it requires careful attention to real-world impacts, particularly on those whose voices are least likely to be heard.

By focusing on AI governance with the charitable sector, this White Paper addresses an important gap. Charities hold public trust and exercise significant influence over individuals and communities experiencing disadvantage. Ensuring that their use of AI sends clear signals of trustworthiness is not only good governance – it is a human rights imperative.

Australia has the opportunity to embed human rights into the governance of AI in ways that are practical, proportionate and forward-looking. This is especially important in the charitable sector, where innovation must always serve the public good and where the rights and dignity of those experiencing structural disadvantage must remain at the heart of institutional decision-making.

This White Paper is a timely and valuable contribution to the national conversation on responsible, human-centred AI. Dr Brydon Wang and Dr Kim Weinert should be congratulated for their thoughtful and impactful work in an area of increasing urgency and importance.



Executive Summary

Artificial intelligence is rapidly entering the Australian charitable sector. Its adoption is reshaping how charities allocate resources, interact with end-users and beneficiaries, and exercise institutional discretion. This White Paper examines the governance implications of this shift and proposes a trustworthiness-based framework to ensure AI deployment strengthens, rather than undermines, charitable purpose and public trust.

This White Paper makes the following findings:

- AI adoption in charities is driven by structural necessity, not technological ambition. Rising demand, constrained funding, workforce and volunteer fatigue, as well as increasing expectations of efficiency, are compelling charities to adopt AI systems to sustain operations.
- AI is no longer confined to administrative assistance. Charities are using AI in fundraising, communications, compliance and increasingly in the triage and prioritisation of service delivery. These contexts are where algorithmic systems participate directly in decisions affecting access to support.
- Existing governance frameworks are not designed for algorithmic decision-making. Australian charity governance relies on fiduciary duties developed for human judgement. These frameworks assume decision-makers can understand, explain and take responsibility for institutional decisions. AI systems challenge these assumptions by operating at scale while relying on opaque models, distributing responsibility across vendors and data infrastructures.
- AI introduces new risks of accountability gaps and discrimination. Algorithmic systems may replicate or amplify existing structural inequalities, particularly affecting vulnerable individuals who have limited capacity to challenge decisions or seek alternative support.
- AI governance in the charitable sector is fundamentally a question of institutional trustworthiness. Charities exercise power over individuals who are often dependent on their services. The legitimacy of AI deployment depends not on efficiency gains alone, but on whether charities can justify exposing beneficiaries to the risks AI introduces.
- Trustworthiness is distinct from trust. Trust describes the willingness of beneficiaries and the public to rely on charities, often under conditions of vulnerability. Trustworthiness describes the institutional qualities that justify that reliance. Charities cannot assume they are trusted or that the AI systems they deploy are trusted. Instead, charities must demonstrate trustworthiness before deploying AI systems that structure institutional discretion.



This White Paper proposes a Trustworthy AI framework grounded in the three signals of trustworthiness:

- **Benevolence:** AI systems must demonstrably serve the interests of end-users and beneficiaries, not merely organisational efficiency. Without this orientation, the subsequent signals of integrity and ability risk collapsing into procedural compliance;
- **Integrity:** AI use must align with charitable purpose, satisfy the law and human rights obligations not to discriminate, and be transparent and contestable; and
- **Ability:** Charities must possess the institutional capacity to understand, oversee, and where necessary, restrain the AI systems they deploy and rely upon. Where this capacity is absent, restraint is the responsible choice.

This White Paper further articulates six governance principles for charitable AI adoption:

- Charitable purpose must retain its primacy;
- End-user and beneficiary voices must be centred;
- Algorithmic transparency and explainability;
- Proactive bias identification and mitigation;
- Meaningful human oversight and override; and
- Clear accountability allocation.

The central conclusion of this White Paper is that AI adoption by charities is not merely a technical or operational decision. It is a governance decision that affects human rights, public benefit and the legitimacy of charitable purpose itself. The question is not whether charities can use AI but whether they are justified in doing so while remaining trustworthy.

Authors



Dr Kim D Weinert is a leading expert in charity law, governance and regulation, with a particular focus on the legal frameworks that shape the operation, accountability and public benefit obligations of charitable and not-for-profit organisations. She is a Lecturer at the TC Beirne School of Law at the University of Queensland, a member of the Queensland Law Society's Not-for-Profit Committee, a member of Alliance for Social Impact at UQ, charity law scholar with the International Charity Law Network, and Visiting Fellow (2025-2026) at the State Library of NSW.

Kim recently co-edited *Charity Law and Governance: Private Benefit, Public Benefit and the Regulatory Strategy* (Hart Publishing, 2025), which explored the need for legal frameworks to evolve in light of changes at the intersection of regulatory compliance, public benefit and organisational sustainability. Kim's research has been cited by the Queensland Supreme Court and the Parliament of Western Australia, reflecting its relevance to both judicial reasoning and legislative reform. Her scholarship is attentive to the practical operation of charity law and governance standards, while also engaging critically with their underlying assumptions and limitations.



Dr Brydon Timothy Wang is a leading expert in the regulation of artificial intelligence, automated decision-making systems and data-driven technologies, with a particular focus on the role trustworthiness plays in infrastructure governance and public accountability. Brydon is an Adjunct Associate Professor at the University of Queensland's Centre for Policy Futures, where his work examines how law and regulation respond to technological systems that exercise power over vulnerable individuals and communities.

Brydon's research centres on the conditions under which institutions are justified in deploying automated systems within the context of infrastructure and public service delivery. He developed a trustworthiness-based framework for assessing automated decision-making, centred on benevolence, integrity and ability, which now underpins his work on AI governance and regulatory design. His research has informed policy development and institutional practice, including a recent briefing to the Queensland Supreme Court Justices on the impact of generative AI on legal practice. Brydon has previously served on the board of a leading community-based not-for-profit organisation, including in the role of Treasurer.





AI, Charities and the New Trust Challenge

AI is no longer a future concern for Australian charities. It is already embedded in everyday operations and used to augment work, such as drafting grant applications, segmenting donors, prioritising service delivery, responding to enquiries and even managing compliance and drafting reports. For many charitable organisations, the adoption of AI systems has been a reflexive response driven by necessity rather than strategic design. At present, there is mounting demand on charities to deliver services and goods to those with increasingly complex needs while constraints are being imposed through funding arrangements set in place through government contracts and grants. This is exacerbated by a decline in volunteerism,¹ higher operational costs and rising expectations from the public, donors and increased regulation. The sum of these pressures requires charities to ‘do a lot more with a lot less’. Consequently, AI systems are seen as the panacea for these challenges.

Charities have always been held out to be more trustworthy than for-profit organisations and the state.² What makes a charity more trustworthy is a question that has been distinctly answered in different ways by the disciplines of economics,³ sociology,⁴ management⁵ and psychology.⁶ However, across disciplines there is consensus that charities are regarded as more trustworthy because they are understood to act in the interests of others rather than their own. How a charity appears and behaves to the public depends on the public judging it for acting in a manner that is good,⁷ selfless, wholesome, altruistic, and benevolent towards those most in need. That is, trustworthiness is signalled through governance practices, such as the stewardship of resources, protection of sensitive information and the exercise of discretion in ways that appear integrous and humane. Where a charity can satisfy and deliver on all these expectations, it sends the requisite signals of trustworthiness, and the public will have confidence and provide support to the charity and the sector.

However, these signals of trustworthiness are increasingly difficult to sustain under conditions of chronic resource constraint. For many charities, the challenge of prioritising limited resources now intersects directly with the need to maintain institutional trust. Organisations are under pressure to invest in software and systems to meet regulatory, funding and operational expectations at the expense of delivering fully on their charitable purpose. At the same time, having to stretch resources means that charities often make difficult discretionary decisions about who receives support, when, and on what terms.⁸ These decisions are unavoidable and are acutely visible to end-users, beneficiaries and the public.

- 1 Nicholas Biddle and Matthew Gray, ‘Ongoing Trends in Volunteering in Australia’ (Australian National University, Centre for Social Research and Methods, 30 October 2023) <https://polis.cass.anu.edu.au/files/docs/2025/6/Ongoing_trends_in_volunteering_in_Australia.pdf>.
- 2 Henry Hansmann, ‘The Role of Nonprofit Enterprise’ (1980) 89(5) *Yale Law Journal* 835.
- 3 Susan Rose-Ackerman, ‘Altruism, Nonprofits and Economic Theory’ in Helmut Anheier (ed), *Nonprofit Organizations: Theory, Management, Policy* (Routledge, 2005) 126.
- 4 Robert Putnam, *Bowling Alone: The Collapse and Revival of American Community* (Simon & Schuster, 2000); Roger Lohmann, *The Commons: New Perspectives on Nonprofit Organizations and Voluntary Action* (Jossey-Bass, 1992)
- 5 Helmut K Anheier, *Nonprofit Organizations: Theory, Management, Policy* (Routledge, 2014).
- 6 See in general, Christopher D B Burt, *Managing the Public’s Trust in Non-Profit Organizations* (Springer Nature, 2014).
- 7 See Kim D Weinert, ‘Understanding the Meaning of Charity: The Art of Doublethink’ in Ross Grantham and Kim D Weinert (eds), *Charity law and Governance: Private Purpose, Public Benefit, and the Regulatory Strategy* (Hart, 2025) 101.
- 8 Australian Productivity Commission, *Contribution of the Not-for-Profit Sector* (January 2010) 14 <<https://assets.pc.gov.au/inquiries/completed/not-for-profit/report/not-for-profit-report.pdf>>.

In this context, turning to data-driven technologies that promise consistency, scalability and data security can appear not only pragmatic but responsible. AI systems that safeguard personal and sensitive information, reduce overreliance on stretched volunteers, while supporting the delivery of services to larger numbers of people, are readily framed as the means of preserving trust even under pressure. However, when charities adopt AI systems, particularly systems that classify, prioritise, predict or exclude, they do not merely introduce a new tool. Instead, these charities are reshaping their discretionary power over communities and vulnerable people. This creates a growing tension in the sector. On one hand, AI promises efficiency, scale and sustainability in an increasingly competitive attention economy. On the other, it raises profound governance questions: Who is accountable when an algorithm denies assistance? How are bias and discrimination detected in systems that few Responsible Persons fully understand? What happens to public trust in the charity when decisions affecting people's lives are automated, opaque or left unchallenged?

Much of the current guidance on AI offered to charities assumes these questions can be managed by extending existing governance frameworks. Responsible Persons, under the Australian Charities and Not-for-Profits Commission's (ACNC) regime,⁹ are encouraged to apply traditional duties of care, diligence and best interests to AI systems, as if algorithmic decision-making were merely another operational risk.

What Needs to Change

This White Paper argues that this assumption is flawed and that existing governance assumptions are no longer sufficient. AI does not simply introduce new tools into charitable organisations. It alters how institutional discretion is exercised. While some AI systems assist human judgement by improving administrative efficiency, this White Paper is concerned with the AI systems that participate directly in eligibility assessment, prioritisation and triage of services delivered to end-users and beneficiaries. In these contexts, AI systems play the role of structuring institutional decision-making.

This shift exposes a governance problem not contemplated by traditional fiduciary models. Charity governance frameworks assume that institutional discretion is exercised by identifiable human decision-makers who can understand, explain and take responsibility for their decisions. AI systems challenge this assumption as their internal reasoning may not be fully visible or contestable and, exacerbating this problem, their outputs may shape institutional action at scale. Further, the operation of these AI systems may depend on external vendors, training data and technical infrastructures beyond the charity's direct control. As a result, institutional authority may be exercised in ways that remain formally authorised but are not meaningfully governed.

The central governance question is therefore not whether AI systems function reliably, but whether charities remain justified in exercising power through them. Where institutional discretion is delegated to automated systems, governance must address whether those systems operate in ways consistent with charitable purpose, human rights obligations and the trust placed in charities by the public and the communities they serve.

This White Paper examines how AI is reshaping the exercise of institutional power within the charitable sector and identifies the resulting governance gap. It demonstrates why existing fiduciary and regulatory frameworks are insufficient to address algorithmic decision-making, and why this creates heightened risks of discrimination, accountability failure and erosion of public trust. In response, it proposes a trustworthiness-based framework grounded in **benevolence**, **integrity**, and **ability**, and articulates practical governance principles to guide charities, regulators and funders in determining when AI deployment is justified and how it must be governed.

9 See 'The Governance Blind Spot' below for an explanation of a Responsible Person of a registered entity.

Why AI Has Arrived in the Australian Charity Sector

AI has entered the Australian charity sector not through a single strategic decision, but through a series of incremental, often pragmatic choices. Charities have begun using AI tools that promise immediate relief from long-standing structural pressures that will free up limited funding, efficiently meet the increasing demand for services, reporting obligations and to combat workforce and volunteer fatigue.

Across the sector, AI is already being used in three broad areas:

- **Operational efficiency**, including document summarisation, financial analysis, policy drafting and internal and external reporting. AI is also being applied to other internal operations, such as recruitment, volunteer management and customer-facing AI chatbots that handle routine enquiries and knowledge sharing. These are tasks that traditionally consume significant staff time but are rarely mission-defining.
- **Fundraising and communications**, charities are deploying generative tools to draft grant applications, tailor donor communications and manage social media engagement. Predictive analytics are increasingly used to identify prospective donors, optimise campaign timing and segment supporter bases.
- **Service delivery and triage**, where charities working in education, health, housing, family services and community support are exploring or deploying systems that assist with risk assessment, prioritisation of clients and allocation of scant resources.

These developments are not occurring in isolation. Charities operate within an increasingly competitive and information-dense environment. Donors, governments and regulators expect data-driven reporting. Donors are accustomed to personalised digital engagement. Governments increasingly frame innovation and efficiency as markers of organisational credibility and efficiency. Against the backdrop of these demands AI adoption can feel less like a choice and more like a condition of remaining visible, relevant and fundable.

There is also a comparative dynamic at play. Charities do not benchmark themselves only against other charities. They observe the rapid uptake of AI across the public and private sectors and worry about how the Australian and New Zealand charity sectors are ill-equipped to respond to AI.¹⁰ This concern creates subtle pressure on charities to adopt AI tools, even when the organisational purpose, risk profile and ethical obligations are markedly different. Importantly, the AI adoption is often accompanied by unease. Many charity leaders express concern about the loss of the personal and relational qualities that define charitable work.¹¹ Others worry about

10 See Infoxchange, 'Digital Technology in the Not-for-Profit Sector Report' (November 2025) <https://www.infoxchange.org/sites/default/files/infoxchanges_2025_digital_technology_in_the_not-for-profit_sector_report_0.pdf>.

11 Angela Aristidou, Andrew Dunckelman and Sam Fankuchen, 'How AI Can Deepen Nonprofit Relationships', *Stanford Social Innovation Review* (Fall 2025) <<https://ssir.org/articles/entry/artificial-intelligence-donor-engagement>>.

embedding systems they do not fully understand or cannot easily explain to boards, staff or beneficiaries.¹² These anxieties coexist with genuine optimism about what AI might enable. The result is a sector that is both curious and cautious, hopeful and uneasy with moving forward without a settled governance compass.

Understanding why AI has arrived in the charity sector is therefore essential. Without recognising the structural pressures, attention dynamics and capacity constraints driving adoption, governance debates risk sounding abstract or moralistic. Charities are not adopting AI recklessly. Instead, they are responding to real challenges. As such, the task is not to dismiss AI's appeal, but to examine whether the ways in which it is being adopted genuinely align with charitable purpose, public benefit and trust on which the sector depends.

The Promise and the Pressure of AI in Charitable Work

For charities, the appeal of AI stems from a practical need that is neither abstract nor ideological. AI tools promise to alleviate bottlenecks that have long constrained charities' efficacy by freeing staff and volunteer capacity from basic (and increasing) administrative tasks so that they can focus on emerging and increasingly complex social needs and issues that are aligned to the charity's purpose. In this sense, AI is often framed as an enabler, a way to sustain charitable activity under conditions of constraint.

The most frequently cited benefits relate to efficiency and scale. Generative AI tools can assist with drafting grant applications, reports, and communications in a fraction of the time required previously.¹³ Data analytics can help charities identify patterns in donor behaviour, program uptake or service demand that would otherwise remain invisible. Automation can reduce repetitive administrative tasks, thereby freeing staff to focus on direct engagement with end-users and beneficiaries. For charities operating with small teams and limited budgets, these gains are not trivial.

AI is also promoted as a means of consistency and objectivity. In contexts where staff turnover is high, or demand exceeds capacity, algorithmic tools are seen as a way to apply uniform criteria, reduce individual discretion and manage risks.¹⁴ Predictive systems promise to help charities target support more effectively by orienting on those identified as most in need or more likely to benefit from intervention. For donors, governments and regulators that are increasingly focused on measurable outcomes, such tools appear to offer greater accountability and evidentiary rigour.

At the same time, AI adoption is shaped by a less explicit but powerful pressure: the need to remain visible and credible in an increasingly competitive information and data environment. Charities now operate alongside government agencies and corporate entities that are rapidly embedding AI into their operations. Digital sophistication is often treated as a proxy for organisational competence. If charities decide not to adopt AI, the decision can be perceived by donors and the government as a failure to innovate or keep pace.¹⁵ This creates a form of ambient pressure rather than an explicit mandate. Charities may feel compelled to experiment with AI tools not because they are convinced of their suitability, but as a reflexive response to avoid 'paralysis by analysis'. Over time, experimentation can lead to reliance, par-

12 Matthew Schulz, 'Charity AI Skills Gap a Global Threat', Institute of Community Directors Australia (18 September 2025) <<https://www.communitydirectors.com.au/articles/charity-ai-skills-gap-a-global-threat>>.

13 Infoxchange, *Digital Technology in the Not-for-profit Sector Report* (November 2025) <https://www.infoxchange.org/sites/default/files/infoxchanges_2025_digital_technology_in_the_not-for-profit_sector_report_0.pdf>; Institute of Community Directors Australia, 'Writing Funding Applications with AI: Opportunities and Pitfalls' <<https://www.communitydirectors.com.au/help-sheets/writing-funding-applications-with-ai-opportunities-and-pitfalls>>;

14 European Parliament, 'Digitalisation, Artificial Intelligence and Algorithmic Management in the Workplace; Shaping the Future of Work: Cost of Non-Europe' (October 2025) III-IV <[https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774670/EPRS_STU\(2025\)774670_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774670/EPRS_STU(2025)774670_EN.pdf)>.

15 See Australian Government, Digital Transformation Agency, 'Accelerating AI in the Public Sector' (Speech, 13 August 2025) <<https://www.dta.gov.au/articles/speech-accelerating-ai-adoption-public-sector>>.

ticularly where systems become embedded in core functions such as fundraising pipelines, client intake or internal decision-making processes.

Yet the very features that make AI attractive also generate unease. Many charity leaders express concern about becoming dependent on systems they do not fully understand,¹⁶ supplied by vendors they cannot meaningfully interrogate, and trained on data that does not reflect the communities they serve. Others worry about the gradual erosion of relational, human-centred practices that are central to charitable work but difficult to quantify or automate.¹⁷

The tension between promise and pressure is critical. AI is not being adopted in a neutral governance environment. Instead, AI systems are being introduced into organisations whose legitimacy depends on trust, whose end-users and beneficiaries often lack meaningful choice or voice, and whose decisions can have profound consequences for people already experiencing vulnerability and marginalisation. The challenge, therefore, is not whether AI can deliver efficiencies, but whether those efficiencies are being pursued in ways that remain consistent with charitable purpose, public benefit and accountability.

The next section examines the current approaches Australian charities are taking in deploying AI systems within their organisations.

What Charities Are Actually Doing

A 2025 survey of not-for-profit and charitable AI adoption reveals the breadth of current deployment.¹⁸ The Australian Red Cross Lifeblood service, for instance, partners with a technology firm, DiUS,¹⁹ to use AI to assist staff in managing complex documents, including research papers, manuals and financial reports. The AI scans materials to generate plain-language summaries with source links, enabling staff to understand issues more quickly while maintaining accuracy.²⁰ Indeed, the use of generative AI is not confined to individual charities but also extends to regulators. The ACNC itself is exploring the use of Microsoft Copilot to streamline operations, conduct financial analysis and power its chatbots, while also trialling AI to extract financial transactions from PDF documents.²¹ Likewise, charities are using generative AI to draft policies, create meeting minutes through speech-to-text tools, summarise dense regulatory guidance and automate routine compliance tasks. The Alcohol and Drug Foundation operates 'Dib', a chatbot providing information about alcohol and drugs by searching the foundation's website for relevant documents.²²

In fundraising, AI applications have become particularly widespread. By 2024, it was reported that 58 per cent of charities had incorporated AI into their communications and 68 per cent were using it for data analysis.²³ Generative AI tools have been

16 Infoxchange, *Digital Technology in the Not-for-profit Sector Report* (November 2025) <https://www.infoxchange.org/sites/default/files/infoxchanges_2025_digital_technology_in_the_not-for-profit_sector_report_0.pdf>; Lauri Goldkind, Joy Ming and Alix Fink, 'AI in the Nonprofit Human Services; Distinguishing Between Hype, Harm and Hope' (2025) 49 *Human Service Organizations: Management, Leadership & Governance* 225.

17 See generally, Nathan Chappell and Scott Rosenkrans, *Nonprofit AI: A Comprehensive Guide to Implementing Artificial Intelligence for Social Good* (John Wiley & Sons, 2025); see also Liang Echo Shang and Shirley Qian Jin, 'A Compassion Paradox: Can AI Truly Bridge the Empathy Gap in Human Relationships?' in *Artificial Intelligence and the Future of Human Relations* (Springer, 2025) 149–171.

18 Matthew Schulz, 'Forty-Three Examples of NFPs and Charities Using Artificial Intelligence', *SmartyGrants* (21 October 2025) <<https://www.smartygrants.com.au/articles/forty-three-examples-of-nfps-and-charities-using-artificial-intelligence>>.

19 See DiUS, 'Australian Red Cross Lifeblood: Boosting Productivity with Generative AI' (Case Study) <<https://dius.com.au/australian-red-cross-lifeblood-generative-ai-productivity-boost/>>.

20 Ibid.

21 Australian Charities and Not-for-Profits Commission, 'Use of Artificial Intelligence (Transparency Statement)' <<https://www.acnc.gov.au/about/corporate-information/our-vision-mission-and-values/use-artificial-intelligence-transparency-statement>>.

22 Matthew Schulz, 'Forty-Three Examples of NFPs and Charities Using Artificial Intelligence', *SmartyGrants* (21 October 2025) <<https://www.smartygrants.com.au/articles/forty-three-examples-of-nfps-and-charities-using-artificial-intelligence>>.

23 DonorSearch, 'AI for Nonprofits: Everything Your Org Needs to Know' (20 December 2024) <<https://www.donorsearch.net/resources/ai-for-nonprofits/>>.

used by charities to draft grant applications, create social media content and personalise donor outreach.²⁴ Predictive analytics identify prospective major donors and optimise fundraising campaigns based on historical giving patterns.²⁵ Some organisations employ AI-powered chatbots to handle routine donor inquiries, freeing staff to manage and undertake more complex relationship management.²⁶

Service delivery applications are expanding rapidly, particularly in social welfare and health charities.²⁷ The Smith Family uses AI, predictive analytics and machine learning to identify children at risk of dropping out of school, targeting early intervention resources toward those deemed most vulnerable.²⁸ Cool.org employs machine learning to generate lesson plans based on known curricula, while Wildlife.AI uses smart camera traps with machine learning to identify species for conservation work.²⁹ Eastern Palliative Care has implemented an AI-assisted biography service that transcribes and helps write life stories of home care clients.³⁰

The Charity Commission of England and Wales (the 'Charity Commission') frames the use of AI by charities optimistically, noting that AI can help charities 'free up valuable time spent on resource-intensive tasks, and so make more hours available for high priority areas.'³¹ Similarly, guidance from Google emphasises that AI tools can 'act as digital collaborators', empowering staff to tackle complex tasks more efficiently, unlock new levels of creativity to make better data-driven decisions. This efficiency narrative is compelling for charities that always seek to do more with less, allowing these organisations to redirect savings toward their mission, while leveraging these technologies to amplify their impact.³²

Enthusiasm for AI-generated content has also created unintended administrative burdens for both regulators and charities. The Charity Commission has received over 1,000 charities registration applications in September in 2025.³³ A significant driver of this surge is the apparent use of AI by applicants to generate responses in application forms for registration. Unfortunately, as the Charity's Commission's Head of Registration observes 'AI-generated content is often too generic and fails to reflect the specific activities or aims of the organisation applying to become a registered charity, leading to a higher rate of rejections.'³⁴ Further, the Charity Commission's CEO has expressed concern that as AI improves, identifying AI-generated applications will become increasingly difficult, placing additional strain on already stretched regulatory resources.³⁵ Paradoxically, tools promoted as reducing effort for applicants have increased the administrative burden on regulators, who must now assess and triage growing volumes of AI-assisted submissions that vary widely in relevance, structure and evidentiary quality, a phenomenon often described as 'AI slop'.

Charities face similar pressure in the grants and funding context. Consultants in England and Australia have observed that the use of generative AI is likely to significantly increase the volume of grant applications, as it becomes easier to pro-

24 Philip Schmitz, 'Nonprofit AI: 5 Revolutionary Tools + How to Implement Them', *CharityEngine* (29 September 2025) <<https://blog.charityengine.net/nonprofit-ai>>.

25 Ibid.

26 *WomBot* is a chatbot used by the Wombat Housing Support Services <<https://www.wombat.org.au/my-virtual-case-manager>>.

27 For example, *Lookout* is a platform used by charities that offer home care services.

28 The Smith Family, 'Leslie Loble: AI and the Future of Education' <<https://www.thesmithfamily.com.au/media/stories/executive-team/conversations/ai-and-the-future-of-education>>.

29 See Matthew Schulz, 'Forty-Three Examples of NFPs and Charities Using Artificial Intelligence', *SmartyGrants* (21 October 2025) <<https://www.smartygrants.com.au/articles/forty-three-examples-of-nfps-and-charities-using-artificial-intelligence>>.

30 Ibid.

31 Paul Latham, 'Charities and Artificial Intelligence' (2 April 2024) <<https://charitycommission.blog.gov.uk/2024/04/02/charities-and-artificial-intelligence/>>.

32 Google, 'Responsible AI for Nonprofits', *Google for Nonprofits Help* <<https://support.google.com/nonprofits/answer/15231079>>.

33 Emily Moss, 'Commission Considers Using AI as Charity Applications Surge to 1,000 a Month', *Civil Society* (27 August 2025) <<https://www.civilsociety.co.uk/news/commission-considers-using-ai-as-charity-applications-surge-to-1000-a-month.html>>.

34 Charity Commission, 'An Evolving Charity Sector' (26 August 2025) <<https://charitycommission.blog.gov.uk/2025/08/26/an-evolving-charity-sector/>>.

35 Léa Legraien, 'Commission CEO Concerned About AI-Generated Applications Received by Regulator', *Civil Society* (23 January 2025) <<https://www.civilsociety.co.uk/news/commission-ceo-concerned-about-ai-generated-applications-received-by-regulator.html>>.



duce multiple proposals at speed.³⁶ While AI tools may reduce the time required to prepare applications, over-reliance on AI-generated content risks producing submissions that appear ‘impersonal or generic.’³⁷ As Stef Bell, Manager of Australia’s Funding Centre, has noted, AI-assisted applications are often ‘less personalised and specific to the organisation and their projects,’³⁸ underscoring that no automated tool can substitute for an organisation’s own understanding of its mission and purpose.

This dynamic creates a double bind for charities. On the one hand, they are encouraged, and in some cases pressured, to adopt AI tools to realise efficiency gains. On the other, they must invest significant human effort in assessing and sorting through increased volumes of AI-assisted applications submitted by others, eroding the very efficiencies these tools promise. More importantly, this shift risks undermining the relational and trust-based foundations of charitable activity, as interactions become increasingly mediated by generic, AI-generated material rather than grounded in meaningful engagement with a charity’s purpose and work.

36 Matthew Schulz, ‘How will Grantmakers Cope with AI-Generated Applications?’, *SmartyGrants* (1 May 2023) <<https://www.smartygrants.com.au/articles/how-will-grantmakers-cope-with-ai-generated-applications>>.

37 Institute of Community Directors Australia, ‘Writing Funding Applications with AI: Opportunities and Pitfalls’ <<https://www.communitydirectors.com.au/help-sheets/writing-funding-applications-with-ai-opportunities-and-pitfalls>>.

38 Matthew Schulz, ‘How Will Grantmakers Cope with AI-Generated Applications?’, *SmartyGrants* (1 May 2023) <<https://www.smartygrants.com.au/articles/how-will-grantmakers-cope-with-ai-generated-applications>>.

The Governance Blind Spot

Australian charity governance is built on the familiar legal architecture of company directors' legal and fiduciary obligations and duties. Responsible persons of a registered entity³⁹ are required to act with reasonable care and diligence, to act in the best interests of the charity and for its charitable purposes, and to ensure compliance with applicable legal obligations.⁴⁰ These duties are set out in the ACNC Governance Standards, particularly Governance Standard 5, which requires charities' Responsible People (typically trustees of a charitable trust, board members or committee members of an incorporated not-for-profit organisation) to exercise duties comparable to those imposed on company directors.⁴¹ These duties include:

- Acting with reasonable care and diligence in managing the charity's purpose;
- Ensuring the charity complies with legal requirements, including proper management of people's information and data;
- Ensuring appropriate cybersecurity settings;⁴²
- Managing the charity's financial affairs responsibly; and
- Acting honestly and fairly in the best interests of the charity and for its charitable purposes.⁴³

The ACNC's guidance on AI instructs charities to 'consider how their use of AI might impact their governance or registration obligations,' emphasising that Governance Standard 5 applies to AI-related decisions.⁴⁴ In particular, Responsible People must 'ensure they meet the legal requirements of managing people's information and data' and 'ensure their charity's cyber security settings are appropriate.'⁴⁵ The guidance acknowledges risks including privacy issues, discriminatory decisions from biased AI and potential disconnection from charitable purpose, but offers little con-

39 See *Australian Charities and Not-for-profits Commission Act 2012* (Cth) s 205-30 (ACNC Act).

40 See the 'Simplified Outline: how the governance standards work', *ACNC Regulations* reg 45.1.

41 *ACNC Act* s 45-10; *Australian Charities and Not-for-profits Commission Regulation 2013* (Cth) reg 45 (*ACNC Regulations*). See also Kim D Weinert, 'Legal Duties as Part of the Governance Framework for Incorporated Associations: a comparative analysis' (2014) 29(1) *Australian Journal of Corporate Law* 38.

42 In late 2020 the BBC reported that Blackbaud had been hacked compromising bank information and users' passwords of many charities in the United Kingdom. See Leo Kelion, 'Blackbaud: Bank Details and Passwords at Risk in Giant Charities Hack', BBC (1 October 2020) <<https://www.bbc.com/news/technology-54370568>>.

43 Australian Charities and Not-for-Profit Commission, 'Governance Standard 5: Duties of Responsible People' (2024) <<https://www.acnc.gov.au/for-charities/manage-your-charity/governance-hub/governance-standards/governance-standard-5>>; See also, Australian Charities and Not-for-Profit Commission, 'Managing People's Information and Data' <<https://www.acnc.gov.au/tools/guides/managing-peoples-information-and-data>>.

44 Australian Charities and Not-for-Profit Commission, 'Charities and Artificial Intelligence' <<https://www.acnc.gov.au/tools/guides/charities-and-artificial-intelligence>>; Australian Charities and Not-for-Profit Commission, 'Governance Toolkit Cyber Security' <<https://www.acnc.gov.au/for-charities/manage-your-charity/governance-hub/governance-toolkit/governance-toolkit-cyber-security>>; see also Justice Connect, 'Artificial Intelligence: A Guide to Artificial Intelligence (AI) for Community Organisations' (October 2025) <<https://www.nfplaw.org.au/free-resources/privacy-laws/artificial-intelligence-ai>>.

45 Australian Charities and Not-for-Profit Commission, 'Charities and Artificial Intelligence' <<https://www.acnc.gov.au/tools/guides/charities-and-artificial-intelligence>>.

crete direction beyond urging charities to ‘ensure any risks associated with AI use that are specific to their work are properly managed.’⁴⁶

The challenge being met by this White Paper is that AI does not fit neatly within this governance framework. Existing governance arrangements within Australian charity law borrow directors’ duties from company, trust law and equity, where the implicit model is based on human decision-makers exercising judgement about matters they can understand and thus take responsibility for. A decision to be made by the board could involve examining evidence, debating options and making an informed choice. Where such a decision is subsequently proven to be wrong, the board can be held accountable for failing to exercise reasonable care or for ‘mission drift’. However, this approach assumes that decision-makers can understand the decisions being made, explain the basis on which they are made and trace responsibility (and liability) when things go wrong. The Charity Commission stresses in its guidance notes that ‘trustees remain responsible for decision making’,⁴⁷ and that the decision-making process is ‘not delegated to AI or based on AI-generated content alone.’⁴⁸ Trustees are advised to consider whether an internal AI policy would be beneficial and to think about ‘the advantages and risks—and how these would be managed—in the context of [their] trustee duties and charity objectives’.⁴⁹

This governance challenge is made more complex by the differing roles AI systems can play. Where AI systems merely assist internal administrative functions, governance questions resemble those associated with the adoption of other forms of software. However, where AI systems shape prioritisation, triage, eligibility or direct interaction with end-users and beneficiaries, they alter how institutional discretion is exercised. In these contexts, AI systems do not simply improve operational efficiency but instead participate actively in structuring a charity’s response to vulnerability. As systems assume greater autonomy and operate at scale, governance cannot be confined to questions of technical reliability or procurement diligence alone. What is required, instead, is an examination of the legitimacy of delegating discretionary power to automated systems.

However, algorithmic systems challenge each of these assumptions, particularly as AI systems are opaque by design. Even where vendors provide high-level explanations of functionality, the internal logic of machine-learning models may not be interpretable in a meaningful sense. Boards may approve the use of an AI tool without being able to understand how it reaches particular outputs or generated responses. For example, how variables are weighted or how results may change over time as the system is exposed to new data. This creates an immediate tension with duties of care and oversight: it is difficult to govern what cannot be seen or interrogated.

Governance issues are worsened by scale and automation. Algorithmic tools can quickly make or inform hundreds or thousands of decisions. While each decision may seem low risk individually, its combined impact across an entire program can be substantial. Traditional governance differentiates between major strategic decisions, which require board approval, and operational choices, managed at lower levels. However, AI blurs this line, as automated decisions made at lower levels and in large numbers can have systemic consequences without triggering board review.

Responsibility is also diffused across multiple acts, which creates ambiguous lines of accountability and liability.⁵⁰ Charities rarely develop AI systems in-house. Instead, they procure tools from commercial vendors, rely on third-party platforms, or adopt embedded AI functionalities within existing software. As a consequence, decision-making power is dispersed across boards, executives, frontline staff, vendors, data suppliers and technology developers. When an AI-assisted decision produces harm, such as when a charity uses a tool to assist an assessment process and

46 Ibid.

47 Paul Latham, ‘Charities and Artificial Intelligence’, *Charity Commission* (2 April 2024) <<https://charitycommission.blog.gov.uk/2024/04/02/charities-and-artificial-intelligence/>>.

48 Ibid.

49 Ibid.

50 Catherine Brown, Sharon Christensen, Brydon Wang, Amanda Stickley, Trent Candy and Roushi Low, ‘The Impact of Artificial Intelligence on Duty and Standard of Care in Legal Practice: A “Disrupting Moment” for Laws That Protect Consumers from Economic Harm?’ (2025) 6(1) *ANU Journal of Law and Technology* 85.

denies access to a service or misclassifies a client's need, it is often unclear where accountability and liability properly lie.

Existing governance guidance typically responds to this complexity by encouraging charities to extend their existing governance frameworks to AI use, treating algorithmic systems as an incremental operational risk rather than a structural shift in how discretion is exercised. While this approach is understandable, it places a heavy burden on charities that often lack technical expertise, resources or bargaining power. Boards are expected to exercise oversight over systems they may not understand, using contractual assurances and vendor representations as proxies for genuine governance. In practice, this can reduce governance to a form of procedural compliance rather than substantive accountability. Further, by the time these contracts are in place and the AI system is deployed, these compliance and monitoring regimes established through contractual frameworks are often inadequate to address harms that may only be detected later in the deployment lifecycle.

There is also information and knowledge asymmetry at play. AI vendors typically possess far greater technical understanding than the charities purchasing their products. Smaller organisations, particularly those led by volunteer boards, may struggle to ask the right questions during procurement, let alone to monitor system performance over time. The burden of translating general principles into specific AI governance practices rests squarely with charity boards, many of which lack the technical expertise, resources, or time to undertake this translation effectively. Consequently, governance frameworks that assume informed oversight risk becoming performative rather than effective.

Critically, existing governance arrangements are largely purpose- and organisation-centred, rather than beneficiary-centred. Duties are framed around acting in the charity's best interests and advancing its purposes,⁵¹ but they do not require meaningful consideration of how automated systems affect those who depend on charitable services. Where AI systems disadvantage particular groups, governance frameworks may still be formally satisfied if boards have followed reasonable processes, even as substantive harms occur.

This reveals a structural blind spot. Charity governance has traditionally focused on financial stewardship, regulatory compliance and mission alignment. The adoption of AI systems introduces a different category of risk: the exercise of automated power over people, often in contexts of vulnerability where there is low transparency and these individuals have limited avenues for challenge. Existing frameworks do not adequately capture this shift, resulting in a governance lag where charities are encouraged to innovate and modernise and yet the tools provided to govern that innovation have not kept pace. Without more tailored governance approaches, there is a real risk that AI systems will become embedded in charitable operations without sufficient scrutiny, normalising automated decision-making in contexts where human judgement, discretion and relational engagement have long been central.

The next section turns to the consequences of this governance gap. It examines how algorithmic bias can emerge and persist in charitable settings, and why its effects are particularly acute when AI systems are deployed in organisations tasked with serving marginalised and vulnerable communities.

51 Kim D Weinert, 'Legal Duties as Part of the Governance Framework for Incorporated Associations: A Comparative Analysis' (2014) 29(1) *Australian Journal of Corporate Law* 38

Bias, Power and the Charitable Context

Bias and AI are not strange bedfellows. Bias is a well-recognised structural feature of AI systems given that they may be trained on historical data that reflect existing patterns of inequality, or designed by teams whose composition may not reflect affected communities. These AI systems may also be deployed in contexts where power asymmetries already shape who receives support and on what terms. As previously mentioned, when charities adopt AI systems, they do not introduce neutral tools into neutral spaces. Instead, they embed algorithmic decision-making into relationships fundamentally marked by vulnerability, dependence and a lack of choice and autonomy.

Unlike commercial relationships where consumers can choose alternative providers, those seeking charitable services often cannot. For instance, a person experiencing homelessness cannot shop around when a charity's AI system flags them as 'high risk' and denies them accommodation. A refugee family cannot negotiate when an algorithm determines they are ineligible for settlement support. A parent whose child is identified by predictive analytics as 'at risk of school failure' cannot opt out of that classification. In these contexts, algorithmic decisions do not merely influence outcomes, they can determine access to safety, stability and dignity.

The consequences of biased AI systems are therefore not evenly distributed. They fall disproportionately on communities and individuals already experiencing systemic disadvantage, including First Nations peoples, individuals from culturally and linguistically diverse (CALD) backgrounds, people living with disabilities, those in poverty, members of the LGBTQ community and others whom charitable services are intended to support. When AI systems replicate or amplify existing patterns of discrimination, they do so with the veneer of objectivity and at a scale that exceeds individual human decision-making. This creates a particularly insidious form of structural harm where discrimination is automated and, because it is increasingly difficult to detect or challenge, normalised. While many developers of AI systems acknowledge the risk of bias, and charities may seek to address this through staff awareness and training,⁵² distinctive risks remain in the charitable context where affected individuals have limited capacity to resist or remediate harm.

It is important to recognise that charities frequently serve populations experiencing intersecting forms of vulnerability. Individuals may face discrimination based on race, gender expression and identity, sexual orientation, pregnancy, religion and family responsibility, religious affiliation, political views, their address and housing status, and other protected or marginalised characteristics. Algorithmic systems that fail to account for this intersectionality risk compounding marginalisation through institutional practices intended to be altruistic but that may instead produce oppressive outcomes. The stakes are heightened in the charitable context as services are oriented towards addressing fundamental needs, such as housing, food security, safety from violence, healthcare and education. When AI systems mediate access to these necessities, the consequences are existential rather than inconvenient. A biased customer service algorithm may frustrate consumers and donor expectations. Likewise, a biased housing allocation algorithm may render someone homeless. Accordingly, this asymmetry demands governance standards

52 See Justice Connect, 'Artificial Intelligence: A Guide to Artificial Intelligence (AI) for Community Organisations' (October 2025) <<https://content.nfplaw.org.au/wp-content/uploads/2025/11/Artificial-intelligence-and-your-organisation-guide.pdf>>.

proportionate to the severity of potential harm, yet current practice often applies the same 'light touch' risk management used in low-stakes contexts.

What distinguishes decision-making in charitable contexts is not merely the risk of error, but the difficulty of detecting and correcting those errors once institutional processes have been automated. This is because end-users or beneficiaries are limited by choice unlike consumers in commercial markets who might complain, seek alternative providers or pursue legal action. End-users and beneficiaries may not know that an algorithm was involved in the decision affecting them or may not understand how to request an explanation or review. They may fear that by challenging decisions, the charity may withhold future access to support. Even where they recognise injustice, these end-users or beneficiaries may lack the time, literacy and resources or confidence to navigate complaint processes designed for far more resourced populations. The result is a practical accountability gap in which consequential decisions are made without meaningful opportunities for explanation or redress in what Danielle Citron terms 'technological due process deficits'.⁵³

The rise of algorithmic decision-making in charities is therefore more than just a technological shift. It indicates a shift in how authority is exercised over vulnerable communities. Traditionally, charitable services rely on human interactions (despite their flaws) through face-to-face assessments, caseworker judgements and opportunities where one party might give the other multiple chances for explanation and advocacy. While these relationships were never free from power imbalances, they also allowed for discretion, context-aware decision-making, and human acknowledgment.

In contrast, AI systems reconfigure these dynamics in three key ways:

1. **These systems centralise decision-making power in the hands of those who design, procure and deploy these AI systems.** This creates distance between the decision-maker and the frontline workers who interact with beneficiaries and understand contextual nuances. A caseworker who might exercise discretion to prioritise an applicant facing urgent need may find their judgement overridden by an algorithm that does not recognise urgency outside its predefined parameters. Over time, this can erode professional judgement and transform frontline workers into administrators of algorithmic outputs rather than decision-makers in their own right.
2. **Automation enables decision-making at scale, but this precludes consideration of the beneficiary at an individual level.** An algorithm can process thousands of applications in the time it takes a human to assess one, but this efficiency comes at a cost. The lack of human oversight over the process means that there is the loss of granular attention and human discretion exercised in relation to individual circumstances. For people whose lives do not fit neatly into algorithmic categories, and this disproportionately includes those experiencing marginalisation, there is a risk that automated systems may be systematically unable to recognise their claims to support.
3. **Algorithmic decision-making shifts the burden of proof onto those seeking services.** In human-mediated systems, a person can explain their circumstances, provide context and advocate for themselves. In algorithmic systems, they must fit predetermined categories and provide data in forms the system recognises. Those who cannot navigate these requirements, owing to language barriers, literacy challenges, disability, trauma or simple unfamiliarity with bureaucratic processes, are likely to be excluded not through deliberate discrimination but through structural indifference.

53 Danielle Keats Citron, 'Technological Due Process' (2008) 85 *Washington University Law Review* 1249.

Critically, these shifts in power dynamics can occur without any change in a charity's stated mission or values. An organisation committed to equity, dignity and person-centred care can simultaneously deploy AI systems that undermine those commitments without Responsible Persons being aware of the impersonal reduction of beneficiaries to data points in ways that work against the human-centred values of the charity. This reflects governance failures that allow efficiency considerations to override equity concerns, resulting in a growing gap between how charities describe themselves and how they actually operate algorithmically.

This is where the charitable purpose and public benefit will come under pressure. Charitable status in Australia depends on organisations having a charitable purpose⁵⁴ and satisfying the public benefit test.⁵⁵ When AI systems systematically exclude or disadvantage the very communities or classes of people that charities assert to serve, technology can undermine the public benefit rationale that justifies a charity's privileged legal status, tax exemptions and public trust. The efficiency gains that AI promises, such as processing more applications, reducing administrative costs and scaling services cannot compensate for discrimination against those most in need.

The question, therefore, is not whether AI can make charitable operations more efficient, but whether it can do so while genuinely advancing charitable purposes in ways that respect the dignity, equality and agency of the beneficiaries those charities serve. Where the answer is uncertain, the burden should rest with charities to demonstrate alignment with charitable purpose, not with vulnerable individuals to prove they have been excluded or discriminated against.

Human Rights, Public Benefit and the Fragility of Trust

The governance challenges posed by AI in charitable contexts are not merely technical or administrative. They are, at their core, questions of substantive equality, public benefit and the social compact that underwrites the charitable sector's legitimacy. When algorithmic systems entrench inequality or erode the dignity of intended beneficiaries, these systems affect more than just the operations of the charities and, in fact, interfere with the enjoyment of fundamental rights and undermine the relationship of trust between charities and the communities they exist to serve. Over time, this erosion of trust threatens the public confidence that grants charities their privileged legal and social status. Given the charitable sector's human-centred mission, as well as its uneven historical record in upholding the rights of some segments of the community,⁵⁶ these risks demand careful attention.

AI can lead to discrimination and inequality, unfair treatment of marginalised groups and limiting opportunities based on characteristics such as race or gender.⁵⁷ When this occurs in the charity sector, the damage extends beyond individual harms. The end-users and beneficiaries that charities are supposedly delivering relief then find themselves distrusting the charity to deploy technology fairly. This knowledge then spreads through social networks, deterring others from seeking help. The charity may not realise it has a trust problem because those most affected simply disappear from view.

54 See *ACNC Act* s 25-5; *Charities Act 2013* (Cth) ss 5, 12

55 *Charities Act 2013* (Cth) ss 6-10.

56 Some charities, sometimes called 'discriminatory charities', are legally permitted to discriminate in their operation. Religious organisations and single-sex schools are well-known examples where the law expressly allows exclusionary practice that would otherwise be prohibited. See Adam Parachin, 'What Does it Mean to "Act Charitably?": Revising the Purposes and Activities Distinction in Charity Law' in Daniel Halliday and Matthew Harding (eds), *Charity Law: Exploring the Concept of Public Benefit* (Routledge, 2022); Jane Calderwood Norton, 'Discrimination as Detriment' in Ross Grantham and Kim D Weinert (eds), *Charity Law and Governance: Private Purpose, Public Benefit and the Regulatory Strategy* (Hart, 2025) 123; Jennifer Sigafoos, 'When Should Charities be Allowed to Discriminate? The Case of Single-Sex Services and Transgender People' in John Picton and Jennifer Sigafoos (eds), *Debates in Charity Law* (Hart, 2019).

57 See Australian Human Rights Commission, *Using Artificial Intelligence to Make Decisions: Addressing the Problem of Algorithmic Bias* (Technical Paper, 2020) <https://humanrights.gov.au/__data/assets/file/0025/46465/Final_version_technical_paper_addressing_the_problem_of_algorithmic_bias.pdf>.

Further, the Australian Human Rights Commission's 2021 report, *Human Rights and Technology*, highlighted these concerns, particularly regarding automated decision-making.⁵⁸ The Commission found that AI systems can violate rights to equality and non-discrimination, the right to privacy, the right to freedom of expression, and rights to social security, health and education.⁵⁹ Critically, the report emphasised that human rights protections are not obstacles to innovation but essential safeguards that ensure technology serves privacy needs and human flourishing, rather than entrenching disadvantage.

For charities, this requires reframing AI adoption as a threshold governance decision rather than an operational implementation choice. Before deploying algorithmic systems, charities should (and in some instances do) conduct human rights impact assessments to assess whether and how those systems may affect the dignity and access to services and goods of end-users and beneficiaries. This requires assessing the risk of discrimination, breaches of privacy, unfair prioritisation, potential unconscious bias and other foreseeable impacts on access to charity services, such as housing, healthcare, or education. While there is much more attention and understanding focused on acknowledging and preventing direct discrimination by AI systems, the more insidious and prevalent risk for charities is indirect discrimination.

AI systems can generate discriminatory outcomes even where the criteria they apply appear neutral. The results of these seemingly neutral criteria can have disproportionately adverse effects on people with protected characteristics or attributes.⁶⁰ In the context of charities, this risk is particularly acute where AI systems rely on standardised questionnaires or ostensibly objective criteria to allocate and prioritise resources or determine access to services. When such systems are trained on historical data, they are likely to reproduce existing patterns of bias and disadvantage. Unless charities actively interrogate how apparently neutral algorithmic criteria operate in practice, especially in relation to groups likely to experience indirect discrimination, automated decision-making risks entrenching these disparities instead of correcting them.

Although Australian law at both Commonwealth and State levels prohibits indirect discrimination, the practical enforcement of these protections depends on affected individuals being able to recognise and articulate how discrimination has occurred. Indirect discrimination is inherently more complex and less visible than direct discrimination, particularly where decisions are mediated through opaque or highly technical systems. For many end-users and beneficiaries, it will be difficult, if not impossible to understand how a neutral algorithmic process has disadvantaged them or to identify a basis on which to challenge the outcome. Where these risks of indirect discrimination cannot be adequately identified, explained and then mitigated, AI systems should not be deployed, regardless of any efficiency or administrative benefits they may offer.

Importantly, human rights obligations create affirmative duties beyond prohibitions on high-risk AI deployments. Charities are not just simply required to avoid discrimination but to take proactive steps to ensure AI systems do not perpetuate or amplify existing patterns of disadvantage amongst a charity's current and future end-users and beneficiaries. This includes monitoring for disparate impacts, collecting and analysing disaggregated data where appropriate, acting promptly to address any emerging bias and taking remedial action when discrimination is identified. In some cases, this may require suspending or discontinuing the use of AI systems that cannot be aligned with principles and values of equality.

Charitable status depends on organisations demonstrating that their activities serve

58 See Australian Human Rights Commission, *Human Rights and Technology Final Report: Summary* (2021) <https://humanrights.gov.au/__data/assets/file/0025/46933/Ahrc_rights-tech_2021_final_summary_1.pdf>.

59 Ibid.

60 The difference between the two forms of discrimination is that direct discrimination occurs when a person is treated less favourable owing to a protected characteristic (such as race, gender, or disability), while indirect discrimination occurs when a requirement, condition or practice that appears neutral has a disproportionate adverse effect on people with a protected characteristic and is unreasonable in the circumstances: *Racial Discrimination Act 1975* (Cth) ss 9, 10; *Sex Discrimination Act 1984* (Cth) ss5, 6; *Disability Discrimination Act 1992* (Cth) ss 5, 6.

a public benefit,⁶¹ rather than merely advancing institutional efficiency or sustainability. AI adoption must therefore be assessed on an application-by-application basis that considers each potential adoption by going beyond a consideration of organisational benefits (such as cost savings, administrative efficiency and scalability), to examine if the adoption genuinely advances charitable purposes in ways that serve its end-users and beneficiaries. **This legal test is more stringent than the commercial AI deployment, in which efficiency and profitability provide sufficient justification. For charities, the question is not whether AI makes operations easier, but whether it makes charitable work more effective in advancing its purpose and the public benefit the organisation exists to serve.**

Consider a charity providing legal assistance to people experiencing disadvantage. If it adopts an AI system to triage enquiries, directing simpler matters to automated responses and reserving caseworker time for complex cases, this appears to advance efficiency. But if the AI systematically misclassifies urgent or complex matters submitted by non-English speakers as 'simple' and directs them away from human assistance, the system undermines its charitable purpose and provides little or no public benefit. It may process more enquiries, but it does so in ways that deny meaningful access to those most in need of support. The efficiency gains for the organisation do not constitute a public benefit if it comes at the expense of excluding marginalised groups.

The example above illustrates a fundamental tension in the adoption of AI by charities. Efficiency metrics such as processing inquiries, time saved, and cost reduction are easily quantified and thus attractive to boards and donor support. However, these metrics often fail to capture what matters most in charitable work: the quality of engagement, the appropriateness of support, trust between workers, volunteers, and those they serve (the public), and whether services genuinely meet the needs of marginalised communities and individuals rather than merely processing them through systems. Over time, there is a risk that AI adoption, driven by pressure to demonstrate efficiency and scalability, gradually redefines what counts as charitable success. In this alternate world, efficiency metrics become proxies for public benefit, even when the relationship between them is weak or inverse. Identifying this drift is critical as charitable status is conditional. If charities systematically fail to serve the end-users and beneficiaries they claim to support, or if AI systems exclude those most marginalised, or if efficiency pressures override the quality of care, the public-benefit rationale that justifies their tax-exempt status and other regulatory privileges is weakened. Thus, the charitable sector's social licence depends on delivering genuine benefits that cannot be superseded by organisational interests.

Critically, the obligations must rest with charities to demonstrate that AI systems advance their charitable purposes. Given the power asymmetries inherent in charities' control systems, data and resources, against beneficiaries who often lack the capacity to challenge discriminatory outcomes, proactively establishing alignment with public benefit is the only governance approach consistent with fiduciary obligations. We propose a Trustworthy AI framework for charities to help bring about better alignment with public benefit that focuses on the individual end-user or beneficiary.

61 *Charities Act 2013* (Cth) s 6; *Oppenheim v Tobacco Securities Trust Co Ltd* [1951] AC 297.

Trustworthy AI Framework for Charities

The preceding sections have shown that existing governance frameworks struggle to address the risks posed by AI in the charitable sector. They emphasise compliance, process and organisational interests, but pay insufficient attention to how AI systems exercise power over end-users and beneficiaries and reshape relationships of trust. To address this gap, this White Paper proposes a trustworthiness-based framework for AI governance in charities, grounded in a well-established interdisciplinary model of trust that has been adapted to the distinctive legal, ethical and relational context of the charitable sector.⁶² Rather than treating AI governance as a technical compliance exercise, this framework understands AI adoption as an act of delegated power that must be justified to beneficiaries, regulators and the public.

Importantly, this framework does not treat AI solely as a tool that charities choose to deploy internally. Charities are increasingly operating in environments where end-users, beneficiaries and complainants themselves are using generative AI systems to draft applications, complaints, submissions and responses at scale. This shift is being driven in part by rapid reductions in the cost of accessing capable generative models. The *AI Index Report 2025* highlights that the cost of running models with strong language and reasoning performance has fallen dramatically. For example, the cost of querying a model with GPT-3.5 comparable performance dropped by over '280-fold between November 2022 and October 2024'.⁶³ This cost decline has enabled widespread use of AI for drafting communication, advice and documents by individuals outside of institutional contexts, amplifying procedural volume without corresponding improvements in safeguards or moderation. This interactional pressure brought about by AI is already creating significant administrative burden for charities and oversight bodies alike, particularly as these institutions confront this growing volume of AI-generated material of highly variable quality. In practice, this is forcing decisions about triage, filtering and prioritisation, often with limited resources and under human rights and procedural constraints. Trustworthy AI governance must therefore address not only how charities deploy AI, but how they respond to, manage and structure AI-mediated interaction with the communities they serve.

Trustworthiness is not the same as technical reliability of systems or legal compliance.⁶⁴ It is a governance concept concerned with whether an institution merits the trust placed in it by others. In the context of AI, trustworthiness requires examining not only what systems do, but how they are designed, deployed and overseen within institutional settings where individuals may be exposed to risk without meaningful ability to refuse, contest or exit.⁶⁵ This distinction matters acutely for charities, whose end-users and beneficiaries often engage with services under conditions of dependency rather than choice. As such, in this White Paper, trustworthiness is treated as a threshold condition for deployment, requiring charities to demonstrate

62 Brydon T Wang, 'An Updated Model of Trust and Trustworthiness for the use of Digital Technologies and Artificial Intelligence in City Making' in *Proceedings of the 6th Media Architecture Biennale Conference* (2023) 69–80.

63 Stanford Institute for Human-Centered Artificial Intelligence, *AI Index Report 2025* (2025) 4, 12 and 64 <https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf>.

64 Laina Chan and Brydon Wang, 'Beyond the Black Box: From Hallucination to Proof in Legal AI', Chartered Institute of Arbitrators Australia (4 December 2025) <<https://ciarb.net.au/resource/beyond-the-black-box-from-hallucination-to-proof-in-legal-ai/>>.

65 Brydon T Wang, 'Prompts and Large Language Models: A New Tool for Drafting, Reviewing and Interpreting Contracts?' (2024) 6(2) *Law, Technology and Humans* 88–106, 102 <<https://doi.org/10.5204/lthj.3483>>.

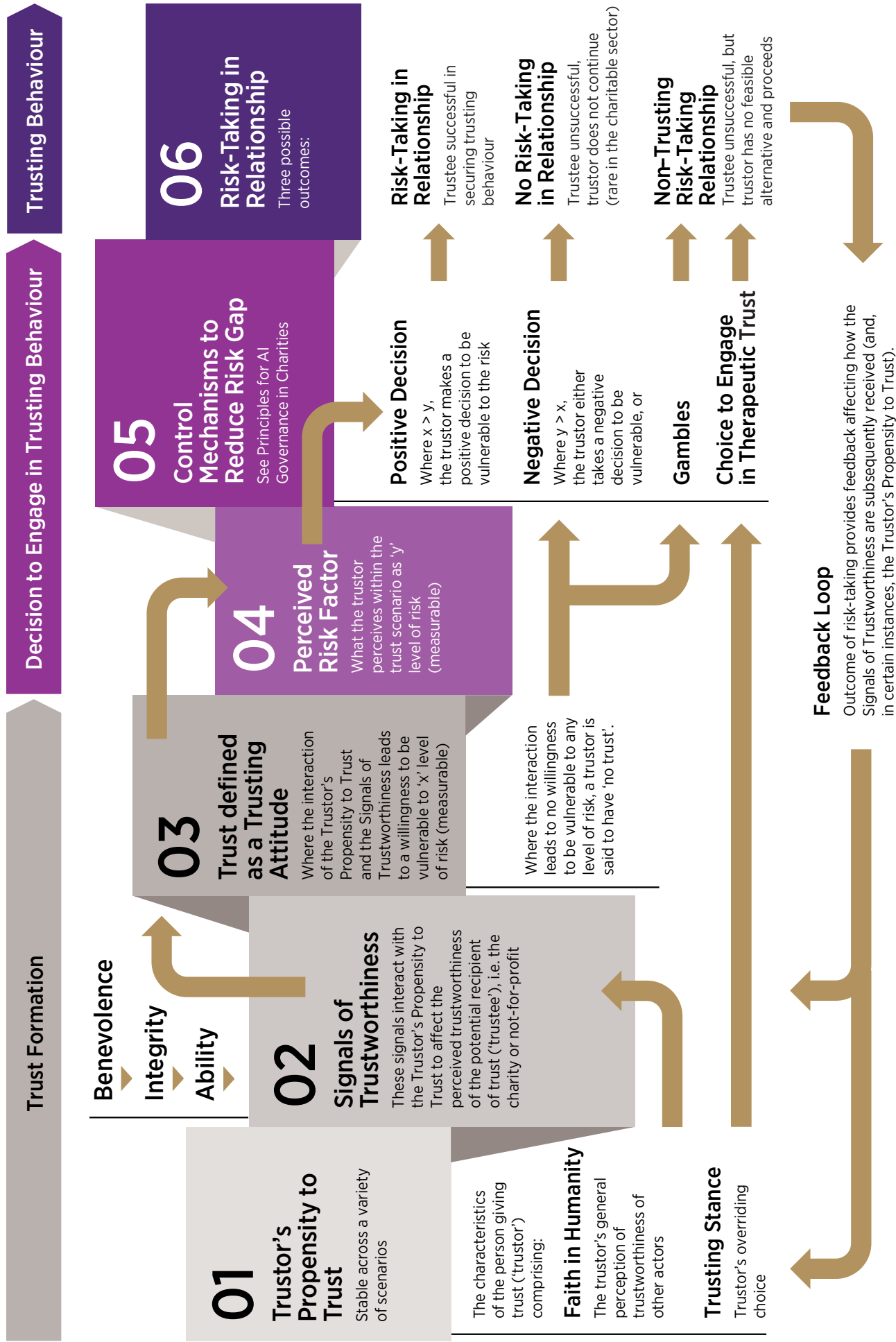


Figure 1: Model of Trust and Trustworthiness

that their use of AI is justified in light of their charitable purpose, governance obligations and accountability to end-users and beneficiaries.

These governance challenges will intensify as AI systems evolve toward increasingly agentic forms capable of initiating actions, managing workflows and interacting autonomously with end-users, beneficiaries and donors. Emerging systems can already conduct conversation, triage enquiries, draft institutional communications and coordinate multi-step tasks without direct human prompting. As these capabilities expand, the boundary between assistive tool and institutional actor becomes increasingly blurred. The governance question, therefore, is no longer limited to whether charities use AI, but how AI systems participate in the exercise of institutional authority itself. This trajectory raises fundamental questions about accountability, contestability and legitimacy.

Figure 1 sets out the trust and trustworthiness model that underpins this framework. The framework set out in this White Paper draws on established trust theory and adapts it to automated decision-making systems. It distinguishes trust as a measurable attitude involving the willingness to accept a certain level of risk from a potential recipient of trust (the 'trustee') as an institutional quality discerned through three interdependent signals of trustworthiness: benevolence, integrity and ability. While all three are necessary, their ordering matters. Without demonstrable orientation toward the interests of beneficiaries (benevolence), procedural compliance (integrity) and technical competence (ability) risk entrenching harm rather than alleviating it.

'Trust' and 'Trustworthiness': Why the Difference Matters for AI in Charities

Trust and trustworthiness are not the same thing.

Trust describes the attitude of a person giving trust (the 'trustor'). In the charity sector, this would be the end-user or beneficiary. Trust can be defined as a willingness to be vulnerable to a particular level of risk. Therefore, given our well-developed ways of describing and measuring risks, the attitude of 'trust' in a trustor is, to a certain extent, measurable against what level of risk they are willing to be exposed to while still continuing to make the decision to trust the charity. However, this willingness to accept the services of a charity might not always be because they want to trust, but because the alternatives are limited or non-existent. In which case, the end-user or beneficiary is taking a gamble to proceed, almost hoping that the charity will prove themselves to be trustworthy.

Trustworthiness, in contrast, describes the qualities of an organisation that justify that trust. This distinction is not merely semantic as the signals of trustworthiness must be observable and are antecedent to the attitude of trust, not a proxy for it. Charities cannot control whether or not an end-user or beneficiary will trust them. This is a complex interaction with the trustor's propensity to trust, their faith in humanity and their trusting stance. What is within a charity's control is how they communicate the signals of trustworthiness, where trustworthiness is a governance obligation and not a reputational asset. It concerns whether a charity has earned the right to ask others to bear risk. This is particularly important when AI systems are used to triage services, assess eligibility, prioritise support or mediate access to care as these are all contexts where errors, bias or opacity can have serious consequences for the vulnerable. Without meeting these trustworthiness requirements, charities run the risk of designing and deploying AI systems that require their end-users and beneficiaries to take a gamble and in doing so, these charities risk the adoption of AI systems that cut across their charitable purpose.

Importantly, this framework does not treat trust as something organisations can create, engineer or optimise. Instead, it treats trustworthiness as a constraint on organisational action, assessed independently of whether trust is in fact placed.

Benevolence: AI Must Demonstrably Serve Beneficiaries, Not Institutions

Benevolence is the most critical and least understood signal of trustworthiness.⁶⁶ It generally refers to the extent to which an institution acts in the interests of those who depend on it,⁶⁷ rather than merely pursuing its own efficiency or sustainability. However, in the context of charities oriented towards beneficiaries and the charitable purpose, this germinal signal of trustworthiness has its clearest articulation. Benevolence in the context of AI deployment looks at whether the charitable organisation can credibly demonstrate that its use of AI is designed to serve the interests of those who bear the risks of algorithmic decision-making, rather than toward organisational convenience or reputational gain. Accordingly, benevolence performs critical ordering work in respect of the other two signals of trustworthiness. Without a demonstrable orientation towards the interests of beneficiaries, the subsequent signal of integrity risks collapsing into formal compliance, and the signal of ability risks being seen as mere technical competence without necessarily being trustworthy.⁶⁸

This is a critical distinction. Many AI systems promise efficiency gains, but efficiency alone may not constitute a public benefit. A charity that deploys AI to process more applications, target more donors or automate the triage of service delivery may improve organisational performance at the cost of disadvantaging particular groups or eroding relational care. Benevolence demands that such trade-offs are not treated as incidental or inevitable. As mentioned above, in the context of charities AI systems frequently operate over people who are experiencing vulnerability: poverty, displacement, disability, trauma or social exclusion. These individuals are not simply end-users of a service, they are often dependent on it and their capacity to refuse, contest, exit and select another provider is limited. In such contexts, benevolence cannot be inferred from intent statements or mission alignment alone. It must be evidenced through governance design.

Perhaps most concerning is that high-risk AI deployment can undermine the relational aspects of charitable work that end-users and beneficiaries value and services they rely upon most. A central concern is the risk of charities 'becoming disconnected from the charity's purpose by losing the personal touch through reduced human connection and empathy.'⁶⁹ This is not merely an abstract concern. Research consistently shows that for individuals experiencing crisis, marginalisation, or trauma, the quality of human relationships with service providers matters enormously to outcomes.⁷⁰

Operationalising benevolence requires charities to ask different questions at the outset of AI adoption. Rather than beginning with 'What can this system do for us?', charities must ask 'Whose interests does this system serve?' and 'Who bears the risk if it fails?' Where end-users and beneficiaries carry the risk of error, bias and even exclusion, the trustworthy signal of benevolence requires that charities practise (and communicate to beneficiaries that it is taking this position of) heightened caution and additional safeguards in examining the design and potential deployment of AI systems.

When charities replace human intake workers with chatbots, automate follow-up through email campaigns generated by AI, or use predictive algorithms to determine who warrants personal contact, they risk treating efficiency as the goal rather than advancing their charitable purpose. This will put a charity's trustworthiness under strain. The Eastern Palliative Care biography service mentioned above exempli-

66 Brydon Wang, 'The Seductive Smart City and the Benevolent Role of Transparency' (2021) 48 *Interaction Design and Architecture(s) Journal IxD&A* 100-121.

67 See Roger C Mayer, James H Davis and F David Schoorman, 'An Integrative Model of Organizational Trust' (1995) 20(3) *Academy of Management Review* 709, 718.

68 Brydon Wang, *The Role of Trustworthiness in Automated Decision-making Systems and the Law* (PhD Thesis, Queensland University of Technology, 2022).

69 Angela Sweeny, Beth Filson, Angela Kennedy, Luci Collinson and Steve Gillard, 'A Paradigm Shift: Relationships in Trauma-Informed Mental Health Services' (2018) 24(5) *BJPsych Advances* 319

70 See Angela Sweeny, Beth Filson, Angela Kennedy, Luci Collinson and Steve Gillard, 'A Paradigm Shift: Relationships in Trauma-Informed Mental Health Services' (2018) 24(5) *BJPsych Advances* 319.

fies this tension. While the assistance provided by the AI system with transcription may indeed help staff capture life stories more efficiently (demonstrating 'ability'), the value of the biography project arguably lies in the human connection involved in storytelling, including the time spent witnessing a life. If AI speeds the process but reduces meaningful human engagement, has the charity actually served its purpose better or simply processed more clients?

This tension between efficiency and relationship is not unique to AI, but algorithmic systems intensify it by making automation seem neutral and objective. When a human staff member decides to spend extra time with a particular client, that judgement draws on empathy, contextual knowledge, and relationship; the qualities that are difficult to quantify but essential to effective care. When an algorithm determines who receives intensive support based on risk scores, the decision appears data-driven and fair, even if the underlying model has no capacity to recognise the non-quantifiable factors that might make the difference between a client thriving or failing.

A charity sends the signal of **benevolence** when its adoption of an AI system:

- prioritises relief over efficiency gains;
- does not shift risk onto beneficiaries least able to absorb it;
- preserves avenues for explanation, discretion and human care;
- provides a process for end-user and beneficiary voices to be heard that seeks to build value consensus;⁷¹
- recognises that doing less with AI may sometimes better serve the charitable purpose than doing more.

Benevolence may also be expressed through practical measures that support meaningful engagement, such as providing plain-language guidance or exemplars to help end-users and beneficiaries understand and use AI tools when interacting with charities, thereby mitigating the procedural overload currently being experienced.

This notion of consensus mechanisms is important to bring together disparate views as part of the overall governance process and to demonstrate benevolence. Trustworthy AI cannot be designed solely by boards, executives or vendors. Charities must instead create channels through which those affected by algorithmic systems can shape their design, raise concerns and challenge outcomes. For example, charities can adopt transparent and participatory approaches to AI governance that allow underlying assumptions, trade-offs and decision pathways to be examined.

These processes create conditions of mutual visibility and mutual vulnerability. In doing so, the organisation does not shield itself behind technical complexity and vendor assurances or claims of inevitability, but allows its choices to be examined by those who bear the consequences of error. By enabling beneficiaries and end-users to understand and contribute to how values are operationalised in AI systems, charities allow community-in-the-loop design principles to flourish and permit legitimate value consensus to form, rather than assuming alignment by default.

⁷¹ Interestingly, the World Economic Forum encourages the conversation about AI philanthropists to have a diversification of voices, but they overlook the end-user and beneficiaries in this conversation. See World Economic Forum, 'Why Philanthropy Need to Prepare Itself for a World Powered by AI' (14 April 2021) <<https://www.weforum.org/stories/2021/04/philanthropists-developed-an-action-plan-for-ethical-ai/>>.

Integrity: Alignment with Charitable Purpose and Existing Laws

The signal of integrity focuses on whether the use of the AI system is consistent with an organisation's stated values, legal obligations and the articulation of its charitable purpose (value alignment). This value congruence, internal to the organisation and outwardly in its service delivery standards and policies, must communicate that this framework of coherence is subject to meaningful accountability. In charitable settings, integrity requires more than technical accuracy or compliance with minimum legal standards. This is because the current approach to applying traditional fiduciary duties to AI governance was designed for a different kind of decision-making that may not be fit for purpose in the context of AI adoption. Instead, AI systems that are designed to signal integrity must demonstrate how the tool is purpose-aligned.

A charity demonstrates **integrity** in its adoption of an AI system when it:

- can clearly articulate how the AI system advances its charitable purpose in substance, not merely in efficiency or scale;
- governs the AI system in a way that is coherent with its legal obligations, particularly in relation to privacy, data protection, charitable law obligations, and its recognition that failure in these domains directly undermines public trust;
- embeds contestability, explanation and review mechanisms so that affected individuals can understand, question and challenge decisions. This allows the charity to ensure that AI-assisted decisions are consistent with the charity's stated values and service standards;
- actively examines and addresses distributional effects, rather than relying on aggregate improvements or average outcomes.

Integrity is therefore not satisfied by improvements in overall performance alone. A system that produces better aggregate results while systematically disadvantaging particular segments of the public undermines integrity, even if it increases efficiency or throughput. Charitable purpose is not satisfied by averages. It requires attention to who bears the burdens of error, exclusion or misclassification, and whether those burdens fall disproportionately on those experiencing vulnerability.

Contestability and explanation are central to this assessment. Where AI systems are treated as inscrutable or unchallengeable, integrity is compromised regardless of technical performance. Beyond this, integrity requires us to consider the deployment of AI systems as a whole within the charity sector. Many charities work with communities already subjected to algorithmic decision-making in government systems. For instance, welfare recipients assessed by automated compliance tools, individuals involved in child protection systems using risk assessment algorithms, and people experiencing homelessness triaged through algorithmic resource allocation. These populations have experienced firsthand how algorithmic systems can perpetuate discrimination, deny benefits based on opaque criteria, and intensify surveillance. Accordingly, to send the signals of integrity to these end-users and beneficiaries, charities need to be alert to how AI adoption may compound, rather than correct, existing patterns of harm.

Integrity demands clear accountability. Charities must retain responsibility for AI-assisted decisions, even where systems are supplied by external vendors. Contractual delegation does not displace governance responsibility. Trustworthy AI requires that accountability remains traceable and that charities are willing to intervene, override, or withdraw systems where harms emerge.

Integrity also requires compliance with legal obligations and relevant standards.⁷² In Australia, the AI governance environment is rapidly thickening, but it remains distributed across cross-sector guidance and existing regulatory regimes, rather than a single, comprehensive AI legislative framework. The Commonwealth has issued cross-sector guidance such as the *Voluntary AI Safety Standard*⁷³ and the *Guidance for AI Adoption*⁷⁴ and announced the Australian AI Safety Institute to strengthen national AI safety settings and work with the National AI Centre from early 2026. Government agencies are also operating under the Digital Transformation Agency's updated *Policy for the Responsible Use of AI in Government*.⁷⁵ At the same time, within the charitable context, expectations are being articulated through the AC-NC's guidance on charities and AI, the privacy and breach-notification framework administered by the OAIC, and other obligations that may flow from international AI safety standards that apply to Australian charities.

These standards, taken together, shape what integrity requires in practice when charities adopt AI and allow these AI systems to ingest sensitive end-user and beneficiary data that it has collected and stored. Compliance with privacy and data protection obligations is particularly salient given the nature of information charities routinely hold, including medical records, financial information and government identifiers. Failures in data stewardship are not abstract risks. In 2023, the data of many Australian charities were compromised when a third-party fundraiser, Pareto Phone, was hacked and data leaked to the dark web. This illustrates how governance failures can propagate through outsourced arrangements. Such incidents underscore that integrity in AI governance extends beyond formal policy adoption to the active management of risk across vendors, systems and organisational capacity.

What this White Paper adds is a centralised organising framework for understanding integrity as a trustworthiness signal under conditions of vulnerability and power asymmetry. While Australian frameworks correctly emphasise transparency, safety and risk management, they can default to a compliance posture that treats trust as something organisations can 'build' through process. In our view, this is a high-risk approach because in the charitable context, end-users and beneficiaries may continue engaging with services not because they trust, but as a consequence of not having meaningful alternatives. Instead, this framework distinguishes trust (the trustor's willingness to accept vulnerability to risk) from trustworthiness (whether the charity has sent out the signals of trustworthiness to earn that right to impose the risk). By sequencing trustworthiness as benevolence first, then integrity, then ability, we insist that integrity cannot be reduced to privacy checklists or cybersecurity controls if the underlying system is oriented towards organisational convenience rather than the interests of end-users and beneficiaries. Only in this way does integrity do the work of constraining AI deployment through demonstrable purpose-alignment, contestability and accountability, while ensuring that legal compliance is treated as a floor rather than the end goal.

Ability: Charities Must Understand and Govern What They Deploy

The signal of ability refers to whether an organisation has the institutional capacity to competently deploy, oversee and, where necessary, restrain the use of AI systems. In the charitable context, this does not require boards or executives to become technical experts. It does, however, require charities to understand, adopt and deploy only governance-capable AI. These are AI systems that can be meaningfully explained, monitored, contested and overridden in practice. Ability therefore speaks to whether a charity is justified in deploying AI at all, given the risks such systems may impose on end-users and beneficiaries.

72 Brydon T Wang and Mark Burdon, 'Automating Trustworthiness in Digital Twins' in Brydon T Wang and CM Wang (eds), *Automating Cities: Design, Construction, Operation and Future Impact* (Springer, 2021) 345–365, 357–359.

73 Department of Industry, Science and Resources, *Voluntary AI Safety Standard: Guiding safe and responsible use of artificial intelligence in Australia* (updated 2 December 2025) <<https://www.industry.gov.au/sites/default/files/2024-09/voluntary-ai-safety-standard.pdf>>.

74 Department of Industry, Science and Resources, *Guidance for AI Adoption* (21 October 2025) <<https://www.industry.gov.au/publications/guidance-for-ai-adoption>>.

75 Digital Transformation Agency, *Policy for the Responsible Use of AI in Government* (ver 2.0, December 2025) <https://www.digital.gov.au/sites/default/files/documents/2025-12/Policy%20for%20the%20responsible%20use%20of%20AI%20in%20Government%202.0_0.pdf>.

A charity demonstrates **ability** in its adoption of an AI system when it:

- understands what the AI system is designed to do, and what it cannot do;
- can explain how decisions are made in terms meaningful to the affected person;
- knows when human judgement must override automated outputs;
- has internal capability to monitor performance, bias and drift over time.

Ability is frequently undermined by procurement practices that treat AI systems as turnkey solutions. Where boards approve AI adoption based primarily on vendor assurances and technical claims or promised efficiency gains, the signal of ability is not being communicated to beneficiaries. In such circumstances, the charity may be able to operate the system but lack the capacity to govern it. A charity that cannot monitor outcomes, detect bias, audit vendor behaviour or respond to errors lacks the ability required for trustworthy AI deployment, regardless of intent or compliance posture.

Ability also requires charities to understand the infrastructure dependencies of AI systems they deploy, including energy and water demands associated with data centres and cloud services, particularly where these have cost, sustainability or service-continuity considerations for publicly funded or resource-constrained environments.⁷⁶ At the same time, ability must be understood in light of organisational resources and governance structure. Charities that rely heavily on volunteers, operate with limited technical capacity or lack dedicated resources for oversight may be particularly vulnerable to system failure, cyber incidents or misuse. These limitations are not merely operational constraints but also affect how a charity can demonstrate the other signals of integrity and benevolence. Without adequate resources, a charity may find it challenging to responsibly steward sensitive personal information and AI-mediated decisions over time, directly affecting how end-users and beneficiaries assess trustworthiness.

Crucially, ability includes the capacity for institutional restraint. Small charities or those operating in high-risk service contexts may reasonably conclude that they lack the capability to govern certain AI systems responsibly. In such cases, deciding not to adopt AI or to delay adoption until governance capacity can be established is not a failure of innovation. It is an expression of sound ability that underpins perception of the charity's trustworthiness.

Trustworthiness as a Governance Threshold

Taken together, benevolence, integrity and ability form a threshold test for the legitimate use of AI in charitable contexts. If any element is absent, AI use may be lawful or efficient, but it is not trustworthy. For charities, this distinction matters as trust cannot be an optional asset or reputational benefit. Instead, trust forms the basis on which charitable status, public confidence and moral authority rest.

The interactional dimension of AI use also sharpens the importance of the signals of trustworthiness articulated above. Where charities are compelled to triage or respond to AI-generated submissions, the risk is not merely inefficiency but exclusion: those with access to AI tools and the skills to use them effectively may dominate institutional attention, while those without such capacity are crowded out. In this context, **benevolence** requires charities to consider whether and how they support end-users and beneficiaries to engage with AI responsibly, including through clear guidance, accessible explanations and exemplars that reduce procedural disad-

⁷⁶ Brydon Wang and Gabriel Wong, 'Can We Power Both Humanity and Machines with Cheap Renewables?', *Renew Economy* (26 March 2025) <<https://reneweconomy.com.au/can-we-power-both-humanity-and-machines-with-cheap-renewables/>>.

vantage. **Integrity** requires that any AI-assisted triage or filtering processes remain contestable, explainable and aligned with charitable purpose rather than administrative convenience. **Ability**, in turn, requires charities to assess whether they have the institutional capacity to manage this emerging cycle of escalation in AI-mediated interaction without normalising exclusion or eroding trust.

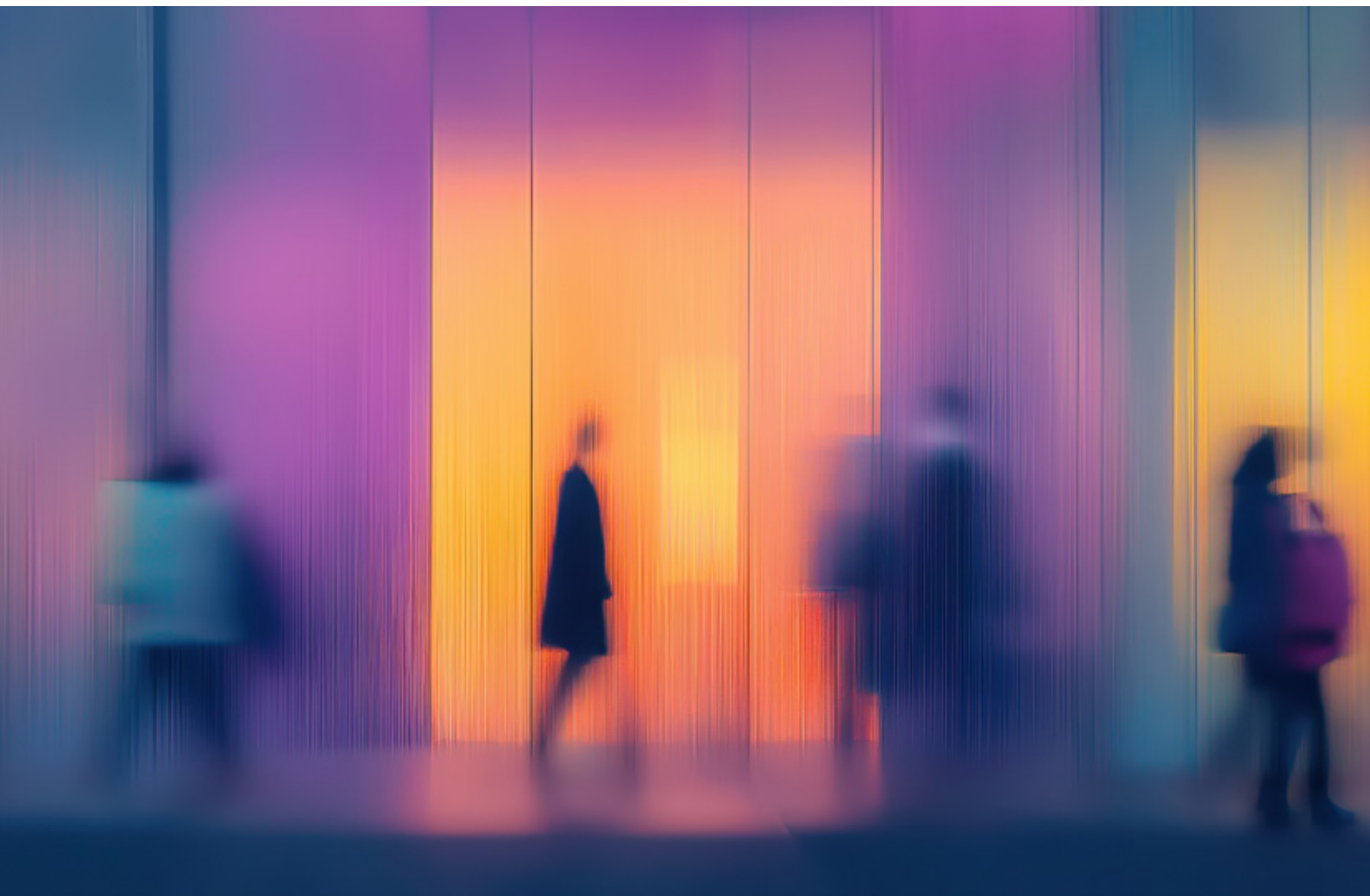
Accordingly, trustworthiness is not something that can be retrofitted through policy statements or post-hoc mitigation. It must be established before deployment, at the point where a charity decides whether it is justified in asking end-users and beneficiaries to accept the risks that AI systems introduce. Thus, the approach taken in this White Paper to examine trustworthiness signals as the evidence of working governance frameworks is not oriented on increasing trust in AI systems, but towards ensuring that organisations are justified in asking others to bear the risks those systems create.

Framed in this way, trustworthiness operates as a governance constraint rather than an aspirational goal. It shifts the central question away from whether AI *can* be used to whether it *should* be used in particular contexts, by particular organisations, and for particular purposes based on the lens of who is bearing the risks of error, exclusion or harm. This framework therefore provides a principled basis for boards to either reject or pause and take time to further consider whether the particular AI system can meet the trustworthiness threshold, even where adoption pressures are strong. The final two sections of this White Paper translate this framework into concrete governance expectations for boards, executives, funders and regulators, recognising that trustworthiness must be supported by institutional practice.



Principles for AI Governance in Charities

The following principles should guide both regulatory frameworks and organisational practices. They build on existing work by the AHRC, international human rights frameworks, and sector-specific guidance, but are tailored to address the distinctive challenges charities face. We note that these principles are intended to function as threshold conditions. Where they cannot be met, charities should refrain from deploying AI systems, even where such systems are lawful, efficient or widely adopted elsewhere.



Principle 1: Charitable Purpose Must Retain Its Primacy



AI adoption must demonstrably serve a charitable purpose. The obligation and responsibility lie with the charity to show how AI deployment advances its mission and benefits those it claims to serve.



Rationale—Efficiency gains do not inherently constitute public benefit. A charity that processes grant applications faster through AI screening has achieved organisational efficiency, but if the AI systematically disadvantages applicants from marginalised backgrounds, it has undermined the charitable purpose. Current frameworks assume efficiency and purpose alignment, but we note that this principle requires demonstrating that alignment rather than merely assuming it.



Application—Before deploying AI, charities must articulate how the system serves its charitable purpose beyond efficiency. If the primary justification is ‘saving staff time’ or ‘reducing costs,’ further analysis is needed: Will saved time be redirected to enhanced services? Will cost savings expand program reach? Or will efficiencies primarily benefit organisational sustainability rather than beneficiaries and end-users?

This principle is particularly important given resource pressures that make efficiency arguments seductive. Charities facing funding constraints naturally seek ways to do more with less. But if ‘more with less’ means processing more clients through algorithmic systems that provide less personalised service or systematically exclude some populations, it may contradict rather than advance charitable purpose. If a charity cannot demonstrate that AI deployment substantively advances its charitable purpose, the system should not be adopted, regardless of efficiency or cost savings.



Practical implication—Responsible Persons should require explicit justification of how proposed AI systems serve a charitable purpose. This should be documented in board minutes and included in annual reports. The ACNC could require disclosure of AI systems’ relationship to charitable purpose in Annual Information Statements, making this analysis a formal governance obligation.

Principle 2: End-User and Beneficiary Voices Must Be Centred



People affected by algorithmic decisions must have a meaningful voice in designing, deploying, and evaluating AI systems. This means going beyond consultation to genuine co-design, particularly for communities most likely to experience algorithmic harm.



Rationale—Current approaches treat AI governance as a technical and legal problem for boards and management to solve. However, those most affected by algorithmic decisions, including end-users and beneficiaries are not involved in the creation and maintenance of governance processes. This reflects a deeper structural problem in charity law, where satisfying the public benefit requirement demands only that a charity’s purposes be directed at a sufficiently broad class of beneficiaries — not that those beneficiaries or end-users have any meaningful participation in, or accountability over, the charity’s governance and decision-making. AI can intensify this problem by creating an additional layer of technical complexity (another barrier) that excludes and oppresses. Consequently, without such mechanisms, claims of benevolence lack legitimacy because value alignment is asserted rather than formed.



Application—Charities deploying AI in service delivery or resource allocation should establish end-user and beneficiary advisory committees with meaningful power to review proposed systems, raise concerns about bias, and recommend alternatives. These committees should adopt community-in-the-loop design principles and include people with lived experience of the issues the charity addresses, and should compensate them appropriately for their expertise and time.

For AI systems affecting vulnerable communities, consultation should occur *before* vendor selection, not after deployment. Beneficiaries and end-users should have a voice in defining what ‘fair’ allocation looks like, what data is appropriate to collect, and what outcomes matter for evaluation. These insights should then be built into the tender documents and technical briefs that vendors will need to address. We suggest that participation of community members should be valued and resourced through providing stipends, accessible venues, interpretation services, and whatever support enables genuine participation.



Practical implication—The ACNC could require charities deploying AI in high-impact contexts (e.g., service allocation, risk assessment, eligibility determination) to demonstrate beneficiary consultation in their governance documents. Funders could make grants conditional on co-design processes that centre on affected communities. Sector guidance could provide models for effective beneficiary engagement in AI governance.

Principle 3: Algorithmic Transparency and Explainability



Charities must be able to explain, in accessible language, how their AI systems make decisions. ‘Black box’ systems incompatible with meaningful accountability should not be used when decisions significantly affect end-users and beneficiaries’ access to services or resources.



Rationale—Accountability requires explanation. If a charity cannot explain why an algorithmic system made a particular decision, it should be held accountable for the failure to ascertain that decision’s fairness or accuracy. Reliance on claims of algorithmic, proprietary, or complexity as grounds for non-disclosure is inconsistent with the transparency and accountability obligations that are foundational to charitable status. These obligations exist precisely because charities operate for the public’s benefit.

This principle extends beyond technical explainability to accessible explainability. This is a shift from a question like: ‘Can data scientists understand the algorithm?’ to the more important question of: ‘Can end-users and beneficiaries understand how decisions affecting them are made?’. A technically accurate explanation involving hundreds of weighted variables is meaningless to someone seeking emergency housing who is told, ‘The system says you’re low priority.’ Meaningful transparency requires translation of algorithmic logic into terms that affected individuals can understand and engage with. Accordingly, AI systems that cannot be meaningfully explained to affected individuals are incompatible with charitable accountability and should not be used in high-impact contexts.



Application—Before deploying AI systems, charities should require vendors to provide both technical documentation (for expert review) and accessible explanations (for end-users and beneficiaries to review and assess). For systems making consequential decisions, charities should be able to provide individuals with meaningful explanations of how their specific cases were assessed using the AI system.

We are not suggesting that this will require proprietary code to be revealed, nor do we intend for our suggestion to create security vulnerabilities. What we argue for is that explaining the factors the system considers, how those factors are weighted, what data is used, and how decisions map to outcomes should be normalised in the charitable sector. If a vendor cannot provide this level of transparency, the charity should not deploy the system for high-stakes decisions.



Practical implication—The ACNC could publish additional guidance on what constitutes adequate transparency for charitable AI, including examples of accessible explanations. Procurement guidance could require contracts to include transparency provisions. Annual reporting could require charities to disclose what AI systems they use and how they ensure explainability.

Principle 4: Proactive Bias Identification and Mitigation



Charities using AI for consequential decisions must conduct regular, independent audits of outcomes disaggregated by demographic characteristics to identify discriminatory patterns. Results should be publicly disclosed (within privacy constraints) and inform system modification or discontinuation.



Rationale—Good intentions do not prevent algorithmic bias. A charity committed to equity can deploy discriminatory AI without realising it may have hidden bias through its training data, design choices and feedback loops invisible to casual observation. Proactive auditing is essential to surface bias before it causes systematic harm. Such auditing must be independent and conducted by those who are not involved in the design or procurement of the system. We note that it is essential that the audit examine actual outcomes, not just algorithmic design. A system that appears fair in testing may produce discriminatory outcomes when deployed in the real world. Auditing must disaggregate results by race, ethnicity, gender, disability, sexuality, age, and other protected characteristics to identify disparate impacts.

Public disclosure (in aggregate, not identifying individuals) serves multiple purposes. It demonstrates accountability to communities or a class of persons that benefits from a charity’s activities. It enables external scrutiny by researchers, advocates and affected communities. It creates reputational incentives for addressing bias. And it contributes to sector-wide learning about what works and what fails.



Application—Charities deploying AI for service allocation, risk assessment, or resource distribution should commit to annual bias audits examining whether outcomes differ across demographic groups. If disparate impacts are identified, the charity should investigate causes to ask: Is it the algorithm, the training data, implementation, or broader systemic factors? The charity should then take corrective action, including (where appropriate) discontinuing use of the system.

For smaller charities lacking resources for independent audits, sector infrastructure could provide shared auditing services. Philanthropic donors could support creation of bias auditing tools tailored to charitable contexts, similar to the AHRC’s human rights impact assessment tools for specific sectors.

We note, however, that the presence of auditing mechanisms does not legitimise AI systems that undermine charitable purpose or exclude particular groups. These mechanisms merely enable earlier detection of harm.



Practical implication—Proposed AI legislation could require charities serving vulnerable populations to conduct bias audits as a condition of using AI for high-stakes decisions. The ACNC could provide templates and guidance for audits. Donors could require audit results as part of grant reporting.

Principle 5: Meaningful Human Oversight and Override



Consequential decisions affecting vulnerable individuals must involve meaningful human oversight, with clear mechanisms for humans to override algorithmic recommendations when judgement, context or equity considerations warrant it.



Rationale—As the ACNC and the Charity Commission emphasise, trustees remain responsible for decisions and must not delegate decision-making to AI alone. Such responsibility requires the capacity to exercise judgement. If frontline staff are expected to ‘oversee’ algorithmic recommendations without authority to override them, or if organisational culture treats algorithmic outputs as authoritative, oversight becomes performative rather than meaningful.

Human oversight is especially vital when AI systems impact vulnerable populations, who may not have the ability to contest decisions. Human judgement can identify situations that algorithms overlook, such as contextual factors, cultural nuances and urgent needs that can invite the application of discretion in ways algorithms cannot.



Application—Charities using AI for service allocation, eligibility determination, or risk assessment should ensure that human decision-makers have authority to override algorithmic recommendations and are trained to exercise that authority appropriately. Decisions that are overridden should be documented, and when a pattern of overrides occurs, this should automatically trigger a review of whether the algorithm requires modification.

Importantly, staff exercising override authority should not face pressure to justify departing from algorithmic recommendations. Organisational culture must treat human judgement as legitimate, not as an inefficient exception to algorithmic optimisation. If the pattern is that staff almost never override the algorithm, oversight may be nominal rather than real. Given this, it is important to note that human oversight is meaningless without recognised authority within the charity.



Practical implication—The ACNC could require charities to document AI override policies and report annually on how frequently humans override algorithmic recommendations. Patterns of very high (suggesting the algorithm isn’t useful) or very low (suggesting staff lack real authority) override rates could trigger further investigation.

Principle 6: Clear Accountability Allocation



Legal and ethical responsibility for AI system outcomes must be clearly allocated to identifiable actors within the charity. This is typically the board for system adoption decisions and executive leadership for implementation and monitoring, rather than diffused across vendors, technology providers and multiple organisational stakeholders.



Rationale—AI creates accountability deficits by distributing responsibility across multiple actors. Clarity about who is accountable is essential to ensuring someone can be held responsible when systems fail or cause harm. Accountability must be allocated in advance of deployment and not constructed after harm occurs.

We note that this principle does not absolve vendors of their responsibilities, but recognises that charities cannot evade accountability by pointing to vendor-supplied systems. When a charity chooses to deploy an AI system, its board and leadership accept responsibility for that choice and its consequences. The fact that a vendor designed the algorithm or that the charity lacked technical expertise to fully understand the system does not eliminate accountability.



Application—Charity boards should designate specific directors (or sub-committees) responsible for AI governance oversight. Executive leadership should designate staff responsible for ongoing monitoring, bias identification, and system evaluation. These accountability assignments should be documented in governance policies and disclosed in annual reports.

When AI systems cause demonstrable harm, such as systematic discrimination, erroneous denials of service, and privacy breaches, designated accountable parties should be required to investigate and report findings, and then implement remediation. The ACNC should have clear authority to investigate and work with other federal government regulatory agencies and charities to investigate and regulate AI-governance failures.



Practical implication—Updated ACNC guidance on Governance Standard 5 could specify that the duty of care includes ensuring AI systems do not perpetuate AI-generated discrimination, with board-level responsibility for system selection and ongoing monitoring. Breaches could be investigated using existing ACNC enforcement powers.

Implementation

Principles without implementation mechanisms remain aspirational. Effective AI governance requires action at multiple levels: within charities, by the sector, through regulation, and via funding structures. At the same time, the charitable sector can reduce individual organisational burdens. For example:

- **Shared Auditing Services:** Rather than each charity conducting independent bias audits, sector infrastructure could provide shared services that can pool expertise, standardised methods, economies of scale. The ACNC, philanthropic donors, or sector peak bodies could sponsor such services.
- **Template Impact Assessments:** Instead of each charity developing bespoke algorithmic impact assessment processes, sector-wide templates could be adapted to specific contexts. The AHRC's human rights impact assessment tools for banking and insurance provide models.
- **Model Policies, Exemplars and Contracts:** Sector guidance could provide model AI policies, exemplars, vendor contract provisions, end-user and beneficiary consultation processes, and override protocols that charities adapt rather than create from scratch.
- **Training and Capacity Building:** Investments in sector-wide AI literacy that includes training and guidance materials for board members, staff, volunteers, end-users and beneficiaries, would build governance capacity more efficiently than each organisation developing expertise independently.
- **Regulatory Safe Harbours:** For small charities implementing AI in limited contexts and following sector guidance rigorously, the ACNC could provide a safe harbour from enforcement action if they demonstrate good-faith compliance with best practices, even if resource constraints prevent gold-standard governance. This would distinguish between under-resourced organisations trying to govern responsibly and those recklessly deploying AI without regard for risks.

Ultimately, however, if rigorous governance proves too resource-intensive, the appropriate response may be for charities to use less AI rather than AI with inadequate governance. Protecting vulnerable populations from algorithmic discrimination must take priority over organisational efficiency.

For Charities: Organisational Practices

Pre-Deployment Requirements

- Conduct algorithmic impact assessments before implementing AI in consequential contexts (service allocation, eligibility, risk assessment, hiring).
- Consult end-users and beneficiaries or their representatives about proposed systems.
- Require vendors to provide both technical documentation and accessible explanations.
- Establish clear policies on human override authority.
- Designate board and staff responsible for AI governance.

Ongoing Obligations

- Annual bias audits examining outcomes across demographic groups.
- Regular review of AI systems' alignment with charitable purpose.
- Staff training on AI limitations, bias risks, and override authority.
- Accessible mechanisms for end-users and beneficiaries to challenge algorithmic decisions.
- Public disclosure of AI use and audit results in annual reports.

End-User and Beneficiary Engagement

- Establish advisory committees, including people with lived experience.
- Compensate (nominally) end-user and beneficiary advisors appropriately for their expertise.
- Design consultation processes accessible to people living with disabilities, people from culturally and linguistically diverse (CALD) backgrounds, and people experiencing literacy barriers.
- Create feedback mechanisms through which end-users and beneficiaries can report concerns about AI systems.

Procurement Standards

- Require transparency provisions in vendor contracts.
- Create procurement pathways that require vendors adhere to a use of AI standard that is trustworthy.
- Demand evidence of bias testing before deployment.
- Include ongoing monitoring and remediation obligations
- Ensure contracts allow independent auditing.
- Negotiate terms that permit system discontinuation if bias is identified.

For the ACNC: Enhanced Regulatory Guidance and Oversight

Updated Guidance on Governance Standard 5

- Specify that reasonable care includes ensuring AI systems do not perpetuate discrimination.
- Provide examples of adequate vs. inadequate AI governance
- Clarify that boards cannot evade responsibility by claiming technical complexity.
- Explain what algorithmic transparency requires in charitable contexts.
- Offer templates for algorithmic impact assessments.

Reporting Requirements

- Require charities to disclose AI use in Annual Information Statements to the ACNC, specifying:
 - What systems are deployed and for what purposes?
 - Were bias audits conducted?
 - How were end-users and beneficiaries consulted?
 - What oversight mechanisms exist?
- Charities serving vulnerable populations should be required to provide more detailed reporting on the above to support an assessment of algorithmic fairness.

Investigation Powers

- Clarify the ACNC authority to investigate AI-related governance failures.
- Develop capacity to assess algorithmic discrimination complaints.
- Partner with technical experts for complex investigations.
- Use investigation findings to develop sector-wide guidance.

Proactive Support

- Develop model AI policies that charities can adapt.
- Provide case studies of good and poor practice.
- Create accessible resources explaining AI risks and governance.
- Offer webinars and training on AI governance obligations.
- Publish plain-language guides.

For Lawmakers: Charity Sector Implementation

Regulatory Priorities

- Do not assume a single AI statute is imminent. Australia does not currently have AI-specific legislation and currently regulates AI through a combination of voluntary standards, regulator guidance and consultation on mandatory guardrails for high-risk applications.
- In this environment, the immediate priority is not charity-specific legislation but ensuring that emerging regulatory frameworks and existing legal regimes remain capable of governing AI systems.
- The ongoing consultations on mandatory guardrails and AI governance create an opportunity to ensure that future regulatory frameworks adequately address the charitable sector.
- As charities exercise institutional discretion over vulnerable populations under conditions of dependency and limited alternatives, any intervention by lawmakers should recognise and respond to this distinctive context.

High-Risk Designation

- Designate AI systems affecting access to charitable services for vulnerable populations as 'high-risk' to ensure that mandatory safeguards are triggered regardless of organisational size or commercial status.
- Define vulnerable populations broadly: people experiencing homelessness, poverty, disability, mental illness, domestic violence, refugees, children in care, First Nations people, members of the LGBTQ community etc.
- Recognise that small-scale deployment affecting vulnerable individuals can be as harmful as large-scale corporate AI systems.

Mandatory Human Rights Impact Assessments

- Require charities deploying AI systems for service allocation, risk assessment, or eligibility determination to conduct human rights impact assessments (HRIAs) using AHRC-developed frameworks.
- Make HRIAs mandatory before deployment instead of treating these as voluntary or 'best practice'.
- Require public disclosure of HRIA results (in aggregate, protecting individual privacy).

Transparency and Explainability Standards

- Require charities using AI for consequential decisions to provide affected individuals with meaningful explanations in accessible language.
- Reinforce that organisations must retain the capacity to review, override and correct AI-augmented decisions.
- Prohibit 'black box' systems where explanation is impossible.
- Set in place a presumption that end-users and beneficiaries have the right to know how decisions affecting them are made.

End-User and Beneficiary Participation Requirements

- For AI affecting vulnerable populations, mandate consultation with representatives of affected communities before deployment, adopting community-in-the-loop design principles.
- Provide funding to support meaningful participation (stipends, accessibility accommodations, etc.).
- Create a presumption that deployment without consultation breaches governance obligations.

Independent Oversight

- Strengthen the capacity of existing regulators, including the Australian Human Rights Commission (AHRC), the Office of the Australian Information Commissioner (OAIC), eSafety Commissioner and Australian Charities and Not-for-profits Commission (ACNC) to address AI-related governance risks.
- Provide clear complaint mechanisms and pathways for individuals harmed by charitable AI.
- Create investigation powers sufficient to assess algorithmic discrimination. While the AHRC has the authority to investigate complaints of unlawful discrimination, its powers are not fully clear to effectively assess and regulate algorithmic discrimination.
- Develop enforcement powers, including system discontinuation orders.

Private Right of Action

- Allow individuals systematically harmed by a charity's AI system to bring civil claims.
- Shift the burden of proof. Once plaintiffs show disparate impact, charities must prove the algorithm serves a legitimate purpose and is the least discriminatory means available.
- Permit class actions for systematic discrimination.
- Ensure remedies include system modification or discontinuation, not merely damages.

For Funders and Philanthropists: Supporting Governance Infrastructure

Change Treatment of Cost of Governance

- Current funding models treat governance as an overhead to minimise rather than as an essential infrastructure for accountability. We advocate that this treatment of the cost of governance must change.

Fund AI Governance, Not Just AI Adoption

- Philanthropic donors supporting AI adoption should ensure funds are allocated for bias audits, impact assessments, and end-user and beneficiary consultation.
- Government contracts requiring service delivery metrics should fund governance capacity to ensure metrics do not drive discriminatory optimisation.
- Multi-year funding should include budget lines for ongoing AI monitoring and evaluation.

Support Sector Infrastructure

- Fund development of shared bias auditing tools tailored to charitable contexts.
- Support creation of algorithmic impact assessment templates.
- Invest in sector-wide capacity building: training, expert networks, accessible guides.
- Sponsor independent research on AI's effects in charitable settings.

Make Governance a Funding Condition

- Require grant recipients to demonstrate AI governance capacity.
- Make funding conditional on conducting bias audits and reporting results.
- Support evaluation of whether AI-enabled programs achieve better outcomes than alternatives.
- Fund long-term studies of AI's impact on charitable purpose achievement.

Create Accountability Mechanisms

- Require reporting on AI governance in grant acquittals.
- Make continued funding contingent on addressing identified bias.
- Publish aggregated findings about sector-wide AI governance challenges.
- Support affected communities to participate in governance oversight.

Conclusion

As charities increasingly adopt AI systems to manage demand, allocate resources and demonstrate impact, they are also reshaping how power is exercised over people who are often experiencing vulnerability and have limited capacity to refuse, contest or exit their interaction with the charity. This White Paper has argued that existing approaches to AI governance are insufficient for the charitable context. Frameworks designed for commercial or governmental settings do not adequately account for the relational, fiduciary and human-centred obligations that define charitable work. In particular, treating AI adoption as a matter of efficiency, innovation or risk management alone obscures the deeper questions at stake as to whose interests are served, who bears the risk of automation, and how decisions that affect people's lives can be made legitimately and accountably.

AI systems span a wide spectrum of use cases, including applications that offer genuine benefits to charities and cannot simply be discarded alongside more problematic deployments. At the same time, charities are not only adopting AI internally but are increasingly receiving AI-generated content from those they serve. This dual exposure introduces a layer of operational and governance complexity that existing frameworks have not yet addressed. As AI-generated submissions and communications increase in volume, charities often lack the resources to meaningfully address them. Over time, this creates a cycle of escalation in which institutional processes become saturated, leading to the normalisation of the risk of exclusion without any actor intending to produce that outcome.

By situating AI governance within a trustworthiness framework grounded in benevolence, integrity and ability, our White Paper offers a way to re-orient decision-making in the charitable sector:

- **Benevolence** foregrounds the interests and dignity of end-users and beneficiaries and performs critical ordering work. Without a demonstrable orientation toward those who bear the risks of AI, integrity and ability risk becoming hollow procedural gestures;
- **Integrity** demands not only alignment between charitable purpose and operational activities, but active engagement with substantive legal principles of discrimination. Specifically, this includes the risk of indirect discrimination, where seemingly neutral algorithmic criteria can result in disproportionate harm to groups already experiencing systemic disadvantage. Charities whose beneficiaries and end-users are subject to algorithmic decision-making in government systems must be alert to the compounding effect of further algorithmic governance, even when each system appears to be legitimate; and
- **Ability** ensures that charities deploy only those systems they can meaningfully understand and govern. This means that a decision not to adopt in inappropriate scenarios should be recognised as a good and responsible governance choice.

Together, these three signals of trustworthiness provide a principled basis for assessing when AI use is justified, when it requires additional safeguards, and when restraint is the most responsible option.

Consequently, the central claim of this White Paper is not that charities should reject the adoption of AI systems. Rather, we contend that AI adoption must be deliberate, transparent and demonstrably aligned with charitable purpose and human rights obligations. Where AI systems undermine equality, erode dignity or displace accountability, they threaten not only individual outcomes but the social licence on which the charitable sector depends. Efficiency gains cannot compensate for the loss of trust that follows when vulnerable communities experience exclusion or harm.

Ultimately, the question facing charities is not whether they can use AI, but whether they can do so in ways that merit continued trust. In the algorithmic age, trust will not be preserved by good intentions or mission statements alone. It is earned through governance choices that place human rights, public benefit, and accountability at the centre of technological decision-making.





CREATE CHANGE

School of Law

Forgan Smith Building
St Lucia Campus
University of Queensland QLD 4072
w law.uq.edu.au

Dr Kim D Weinert

✉ k.weinert@uq.edu.au

Centre for Policy Futures

St Lucia Campus
University of Queensland QLD 4072
w policy-futures.centre.uq.edu.au

Dr Brydon T Wang

✉ brydon.wang@uq.edu.au