Issues Paper Submission
Digital Technology Taskforce

Positioning Australia as a Leader in Digital Economy Regulation
Automated Decision Making and AI Regulation

22 April 2022

Brydon T. Wang
Lecturer
TC Beirne School of Law
University of Queensland
Brydon.Wang@uq.edu.au

Associate Professor Mark Burdon
School of Law / Digital Media Research Centre
Queensland University of Technology
m.burdon@qut.edu.au

TABLE OF CONTENTS

1. Introduction

Thank you for the opportunity to make a submission in response to the *Digital Technology Taskforce* Issues Paper 'Positioning Australia as a leader in digital economy regulation – Automated Decision Making and AI Regulation' (***Issues Paper***). Author details are provided at the end of the submission.

Our submission responds to the Issues Paper that addresses the key objective to build trustworthiness in new technologies and regulatory frameworks that govern new technologies. We see this as the critical pillar under the Commonwealth's *Digital Economy Strategy* to establish the 'right foundations to grow the digital economy'. In our view, many of the questions raised in the Issues Paper should be answered from a position that prioritises trustworthy design and deployment of AI and ADM. By targeting regulation of AI and ADM to ensure these new technologies are perceived as trustworthy by the general public, this overarching regulatory approach will assist to:

- overcome regulatory barriers by ensuring that AI and ADM are seen to be trustworthy, particularly in relation to how AI and ADM developers articulate how they are designing and developing trustworthy applications of technology. Building in regulatory components of trustworthiness at the uptake of new technologies will enhance public acceptance and thus allow the potential for new technologies to be realised whilst maintaining necessary legal requirements;
- build a technologically neutral approach to AI and ADM regulation and will enhance public security and confidence in these new technologies while ensuring that regulation remains legally fit-for-purpose and current;
- identify existing and emerging risks of adopting AI and ADM, particularly a black box approach to using automated decision-making that renders such decision-making challenging to explain; and
- ameliorate the adverse implications of automated decision-making on vulnerable groups.

In our view, a human-centred regulatory approach based on enhancing trustworthiness will position Australia as a leader in digital economy regulation. We make references in the submission to the following publications:

- Brydon Wang and Mark Burdon, 'Augmenting Superintendent Discretion: Trustworthiness and the Automation of Construction Contracts' (2021) 2(1) *Australian National University Journal of Law and Technology* 119 (***Augmenting Superintendent Discretion***)
- 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, Advances in 21st Century Human Settlements, 2021) 345 (***Automating Trustworthiness in Digital Twins***)
- Mark Burdon and Brydon Wang, 'Implementing COVIDSafe: The Role of Trustworthiness and Information Privacy Law' (2021) 3(1) *Law, Technology and Humans* 1 (***Implementing COVIDSafe***)

We summarise our submission in response to the Issues Paper in section 1.2. However, before we get that far it would be useful to briefly outline our concept of trustworthiness as it is integral to understanding the basis for our submission.

1.1 What is trustworthiness?

The concept of trust is difficult to define or measure.[1] Conceptualising trust requires balancing different disciplinary perspectives while ensuring that the framework does not become 'inordinately abstract'.[2] The framework needs to be grounded in everyday practice to avoid the concept of trust becoming too vague.[3]

Our submission adapts the model of trust proposed by Mayer et al (the ***ABI model***)[4] and focuses on the component factors of perceived trustworthiness that a potential recipient of trust (the ***trustee***) might signal to the person giving trust (the ***trustor***). The original model was initially designed to describe the formation and re-formation of trust in organisational settings.[5] However, because the model drew from multiple social disciplines, it has broad appeal and has been widely used across numerous domains of expertise.[6] Relevantly for this submission, the model has been deployed in the area of contract law,[7] automated decision-making,[8] and information privacy. [9]

The model's trustworthiness factors are:

- ***ability***—the demonstration of a particular skill set required in the trust scenario;
- ***integrity***—the trustee's demonstration of compliance with the same set of values and social norms as the trustor; and

---

1 Catholijn M Jonker and Jan Treur, 'Formal Analysis of Models for the Dynamics of Trust Based on Experiences' (Springer, Lecture Notes in Computer Science, 1999), in *Proceedings of the European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW): Multi-Agent System Engineering* 221, 221, citing G Elofson, 'Developing Trust with Intelligent Agents: An Exploratory Study' (1998), in *Proceedings of the First International Workshop on Trust* 125. See also: Thomas W Simpson, 'What Is Trust?' (2012) *Pacific Philosophical Quarterly* 550, 550; Mila Hakanen and Aki Soudunsaari, 'Building Trust in High-Performing Teams' (2012) 2(6) *Technology Innovation Management Review* 38, 38.

2 GA Bigley and JL Pearce, 'Straining for Shared Meaning in Organizational Science: Problems of Trust and Distrust' (1998) 23 *Academy of Management Review* 405, 415.

3 D Harrison McKnight and Norman L Chervany, 'Trust and Distrust Definitions: One Bite at a Time' in R Falcone, M Singh and Y–H Tan (eds), *Trust in Cyber-Societies* (2001) 27, 30, citing C Osigweh, 'Concept Fallibility in Organizational Science' (1989) 14 *Academy of Management Review* 579.

4 Mayer, Roger, James Davis and David Schoorman, 'An Integrative Model of Organizational Trust' (1995) 20(3) *The Academy of Management Review* 715, 715–23.

5 Mayer, Davis and Schoorman, n **Error! Bookmark not defined.**, 711.

6 According to Google Scholar, Mayer et al's article introducing the ABI model has been cited over 26,000 times across a range of disciplines.

7 Mayer et al's article has been cited 3,000 times in the area of contract law, including: Deepak Malhotra and Fabric Luminneau, 'Trust and Collaboration in the Aftermath of Conflict: The Effects of Contract Structure' (2011) 54(5) *Academy of Management Journal* 981, 982; Cristina C Cruz, Luis R Gómez-Mejia and Manuel Becerra, 'Perceptions of Benevolence and the Design of Agency Contracts: CEO-TMT Relationships in Family Firms' (2010) 53(1) *Academy of Management Journal* 69, 70; and Ellen Lau and Steve Rowlinson, 'The Implications of Trust in Relationships in Managing Construction Projects' (2011) 4(4) *International Journal of Managing Projects in Business* 633, 637.

8 Ewart J de Visser, Marvin Cohen, Amos Freedy and Raja Parasuraman, 'A Design Methodology for Trust Cue Calibration in Cognitive Agents' in R Shumaker and S Lackey (eds), *Virtual, Augmented and Mixed Reality: Designing and Developing Virtual and Augmented Environments* (Lecture Notes in Computer Science vol 8525, Springer, 2014); 251, 253; Miriam Höddinghaus, Dominik Sondern and Guido Hertel, 'The Automation of Leadership Functions: Would People Trust Decision Algorithms?' (2021) 116 *Computers in Human Behavior* 106635, 2; Kevin Anthony Hoff and Masoda Bashir, 'Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust' (2015) 57(3) *Human Factors* 407, 409; Balazs Bodo, 'Mediated Trust: A Theoretical Framework to Address the Trustworthiness of Technological Trust Mediators' (2020) *New Media & Society* 1, 13.

9 Mayer et al's article has been cited 9,900 times in articles that contain the term 'information privacy', including: Naresh K Malhotra, Sung S Kim and James Argawal, 'Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model' (2004) 15(4) *Information Systems Research* 336; Heng Xu et al, 'Information Privacy Concerns: Linking Individual Perceptions with Institutional Privacy Assurances' (2011) 12(12) *Journal of the Association for Information Systems* 1; Craig Van Slyke, et al, 'Concern for Information Privacy and Online Consumer Purchasing' (2006) 7(6) *Journal of the Association for Information Systems* 16; Tamara Dinev and Paul Hart, 'An Extended Privacy Calculus Model for E-Commerce Transactions' (2006) 17(1) *Information Systems Research* 61.

- ***benevolence***—the perception of the trustee's intent to look out for a trustor's best interests.

Within the context of law and regulation, there is a blurring of boundaries between ability and integrity. This blurring occurs when a trustee is not perceived as competent because their behaviour does not comply with the legislative requirement to demonstrate value alignment with the law, industry standards and social norms. Importantly, we believe the factor of benevolence has an essential role in allowing an AI or ADM to demonstrate trustworthiness in legal scenarios.

Benevolence is defined as 'the extent to which a trustee is perceived to want to do good to the trustor'.[10] Benevolence thus describes a positive orientation within a trusting relationship as 'loyalty' and 'altruism'[11]. As trusting relationships mature, benevolence as a factor of perceived trustworthiness becomes more 'salient' than the integrity factor to the trust relationship.[12] Benevolence emerges from the desire for people to remain in long-lasting relationships with each other—even in AI and ADM contexts where technological intermediaries mediate these relationships. Benevolence therefore underpins a trustee's positive orientation towards a trustor. Consequently, benevolence takes on a germinal role as an antecedent factor to demonstrations of integrity and ability. As outlined below, our submission argues that benevolence requires a certain form of transparency in order to play a wider role in trust formation that is essential to the development of positive, regulatory relationships involving AI and ADM developers, regulators and the public.

10 Mayer, Davis and Schoorman, n 4, 718.
11 Ibid 719.
12 Ibid 722.

1.2 Summary of Key Points

The following key points outline the basis of our submission.

1. Without a regulatory approach prioritising the trustworthy design and deployment of AI and ADM, the overall regulatory framework will not serve to enhance trust in these technologies and will impact adoption of AI and ADM technologies.
2. Information privacy law is a trustworthiness enhancer rather than a regulatory barrier.
3. Regulation must be enacted from a clear position of benevolence that can only be achieved where there is a specific form of transparency. To demonstrate transparency in a benevolent way, we suggest that transparency must support value consensus, must embrace 'seams' in the automated decision-making process, and must be characterised by mutual vulnerability. Where regulation is underpinned by transparent value consensus, it will be seen as integrous and will promote trust in the technology and increased adoption of AI and ADM.
4. By regulating to enhance trustworthiness in such technologies, regulation will become more technology neutral, so that it will apply more appropriately to AI, ADM and future changes to technology.
5. By regulating to enhance trustworthiness in such technologies, such new regulation or guidance will minimise existing and emerging risks of adopting AI and ADM that seek to take a black box approach to decision-making rendering such processes opaque and challenging to explain.
6. AI and ADM that are not transparent will not be seen as benevolent and will not send sufficiently clear signals of trustworthiness to the general public. Such an approach is not appropriate for establishing trust in these technologies.
7. There are current attempts at making AI more explainable, fair and trustworthy – particularly the framework being adopted in the EU. This should be considered in further detail for adoption in Australia. In our view, it would be desirable to have consistency and interoperability with foreign approaches although the integrative model of trustworthiness we have set out in this submission goes one step beyond this and will position Australia as a leader in digital economy regulation.

We hope that our submission assists the Digital Technology Taskforce. Please do not hesitate to contact the lead author if you have any questions.

2. Responses to Issue Paper Questions

2.1 What are the most significant regulatory barriers to achieving the potential offered by AI and ADM? How can those barriers be overcome?

At present, we believe the role of trustworthiness as an underpinning regulatory facet of AI and ADM is not yet fully understood in Australia. The absence of this understanding is currently not a regulatory barrier in and of itself. However, we believe that without proper consideration in the future, the benefits of AI and ADM are more likely to be fully achieved through an overarching regulatory focus predicated on a clear and identifiable understanding of trustworthiness.

For example, we note that the Issues Paper lists 'construction' as a key area in which AI and ADM are currently being deployed.[13] In our view, the construction industry and construction contracts offer examples of where opportunistic gamesmanship by contracting parties diminish trust amongst parties, thus creating a need for a legally and regulatory supported role of a trusted intermediary, the superintendent. Trusted intermediaries in construction play a key role in ameliorating the disparity of knowledge and expertise between the project owner and the contractor. Similarly, this disparity of knowledge and expertise exists in applications of AI and ADM where individuals subjected to such algorithmic processes and even administrators managing such processes may not fully understand ADM process. Accordingly, the use of AI and ADM to augment the role of the superintendent provides a helpful case study to consider the regulatory development of trustworthiness regulatory activities across the wider economy.

In construction contracts, particularly in standard form construction contracts, the role of the superintendent has become increasingly important. The superintendent is a contractually created trusted intermediary who acts as a go-between and 'quasi-arbiter' for the contracting parties.[14] However, two technological developments occurring within contract administration has required the law to re-examine the role of the trusted intermediary. First, construction trusted intermediaries and their role in administering construction contracts are subject to automation pressures. Existing and emerging technologies—in the form of building information models ('BIMs') and ADM—are increasingly being deployed to automate portions of the contract-administration process. Second, and as a consequence of these automation pressures, the trustworthy decision-maker in a construction contract is gradually being augmented or, in some respects, automated, and this requires an examination of whether an ADM system can fulfil the traditional role of the intermediary: to demonstrate trustworthiness before the law. The augmentation of the superintendent is an example of how AI and ADM disrupt established relationships and reshape consideration of legally defined roles. It also provides a helpful example of how a clearly defined notion of trustworthiness can assist with the ongoing development and application of legal and regulatory frameworks relevant to the automation of crucial business processes.

In effect, the various technological platforms that comprise Australia's digital economy potentially act as intermediaries in the same way as the human superintendent serves to disrupt, modify and shape the signals of trustworthiness in construction contracts and the industry as a whole. Therefore, any regulation relevant to AI and ADM needs to consider how ADMs are deployed to augment or automate the intermediary role. Where there is a failure to fully consider AI and ADM's role as intermediaries of trustworthiness in the overarching regulatory framework, any regulation established may not fully

---

[13] Commonwealth of Australia, Department of the Prime Minister and Cabinet (March 2022) *Positioning Australia as a leader in digital economy regulation - Automated decision making and AI regulation - Issues Paper* 4.

[14] Damien Cremean and Natalie Ozer, LexisNexis, *Halsbury's Laws of Australia* (on 7 March 2018), 65 Building and Construction, 'Certification of Building Contracts' [65–855], citing *Minister Trust Ltd v Traps Tractors Ltd* [1954] 3 All ER 136

convey the necessary signals of trustworthiness to the general public and could consequently make overall acceptance and adoption of AI and ADM more difficult.

These issues of trustworthiness are particularly important regarding the complex privacy issues that arise from the use of AI and ADM for enhanced personalised service provision. Ongoing public sentiments that express continuing concerns about the lack of individual control need to be carefully considered as a core part of trustworthiness regulatory considerations. Our research has demonstrated that information privacy law, which is often perceived in some quarters as a barrier to AI and ADM developments, is in fact the opposite. Due and proper consideration of the complex information privacy issues that arise from AI and ADM is a necessary facet of trustworthiness and leads to more transparent development processes that are more aware and responsive to public privacy concerns.[15]

2.2 Are there opportunities to make regulation more technology neutral, so that it will more apply more appropriately to AI, ADM and future changes to technology?

We support the Issues Paper's view that any regulatory approach to AI and ADM must be made more technology neutral. Currently AI and ADM are defined in many different ways. For example, the term 'automated decision-making systems' is often adopted interchangeably to encompass a broad range of technologies across a spectrum of automation. On the augmentation end of the automation scale, technologies, such as Decision Support Systems assist a human actor to analyse information and better understand the field of choices required in making a decision. In the middle of the spectrum, are technologies that make decisions intent on keeping the 'human in the loop'[16] to varying degrees of involvement. At the full automation end of the spectrum, lie technologies that aim to replace human decision-making with little to no human input. AI and ADM also sit within an ecosystem of other technological solutions that interact with these systems. These include data input components or 'cross modal inputs'[17] from various sensors, such as drones equipped with cameras and machine vision software, 3D scanners and mobile phones. As noted in the preceding section, strengthening information privacy law will enhance trustworthiness as an underpinning regulatory facet and will make an immediate contribution to making such regulation more technology neutral as it operates at the point of human-to-computer interaction. Regardless of any forthcoming technological changes, we do not anticipate that this interaction between what is human and what is automated will shift to the point where trustworthiness and information privacy issues cease to be of significance.

2.3 What regulatory changes could the Commonwealth implement to promote increased adoption of AI and ADM?

In our view, increased adoption of AI and ADM requires specific changes to the regulatory approach undertaken by the Commonwealth to promote the trustworthy design and deployment of AI and ADM. Future regulatory approaches predicated on trustworthiness factors should focus on developing a more distinct strategy based on the factor of benevolence.

Regulation must be enacted from a clear position of benevolence that can only be achieved where there is a specific form of transparency. To demonstrate transparency in a benevolent way, we suggest that transparency must support value consensus, must embrace 'seams' in the automated decision-making

15 *Automating Trustworthiness in Digital Twins* 346-50,358-62; and *Implementing COVIDSafe* 11-2.

16 Monika Zalnieriute, Lyria Bennett Moses and George Williams, 'Automation of Government Decision-Making' (2019) *The Modern Law Review* 82(3), 425, 432.

17 Ibid.

process, and must be characterised by mutual vulnerability. Where regulation is underpinned by transparent value consensus, it will be seen as integrous and will promote trust in the technology and increased adoption of AI and ADM.

A more enhanced benevolent approach could be built on three key themes of transparency:

1. Transparency mechanisms that support value consensus.
2. Transparency mechanisms that build seams into ADM processes.
3. Transparency mechanisms that build on mutual vulnerability requirements.

2.3.1 Transparency mechanisms that support value consensus

Going back to the case study of construction, highlighted above, in the context of private contracting, transparent procurement processes that allow the parties to negotiate and come to a value consensus in a bargain can demonstrate benevolence as a positive orientation towards a contracting party. Such transparency depends on the mutual vulnerability of the negotiating parties. In contrast, where there is no transparency in the procurement process, there is likely not to be value consensus and no consequent meetings of the minds[18] (*consensus ad idem*).

Subsequently, during contract administration, benevolence is demonstrated where the parties act in good faith with fidelity to the value consensus established during the procurement process. The superintendent, as a trusted intermediary, is required to demonstrate benevolence by exhibiting fairness, honesty, reasonableness and cooperation in fidelity to the same value consensus in the bargain. However, as the superintendent role in private contracts is subject to automation pressures, the emergent AI and ADM systems that augment and automate segments of the contract administration process must demonstrate benevolence in scenarios beyond quantitative decision-making. In particular, where the superintendent role requires the exercise of discretion, AI or ADM systems may need to reintroduce human intervention into the system to bring in qualitative considerations that align decision-making with the value consensus in the bargain and to comply with the implied duty of good faith. As noted above, we believe these issues of augmentation and automation will be relevant to many industrial and governmental contexts that involve trusted intermediaries, especially those built upon AI and ADM development.

However, in trustworthy private contracting scenarios, the procurement process is, on its surface, transparent and sets out the negotiating points of each counterparty. All parties and their positions are typically revealed through the procurement process and shared with each other to be captured within the provisions of the contractual terms. As AI and ADM systems are scaled up from the contract to the policy level, benevolence as personal orientation based on achieving value consensus is made complex. This complexity sits within the challenges of imbuing transparency in the design and deployment of AI and ADM. In other words, signalling an intention to ‘do good’ or be benevolent towards contractual parties, and society as a whole, is made more complex because the opaque nature of AI and ADM processes fundamentally disrupts established human and legal relationships and designated roles.

Benevolence can also be demonstrated in the give-and-take of commercial relationships. For example, a contractual party can choose not to exercise a contractual right to damages or rectification, thus signalling a positive orientation towards their counterparty. However, the clearest demonstrations of give-and-take lie in the loosely defined duty to act reasonably and honestly and in the duty to cooperate,

---

18 *Carlill v Carbolic Smoke Ball Company* [1893] 1 QB 256 per Bowen LJ, who held that ‘One cannot doubt that, as an ordinary rule of law, an acceptance of an offer made ought to be notified to the person who makes the offer, in order that the two minds may come together. Unless this is done the two minds may be apart, and there is not that *consensus* which is necessary according to the English law … to make a contract’ (at 269) [emphasis added].

and particularly in the exercise of discretion.[19] While private construction contracts may refer to these duties, what falls within their scope is only partially articulated in case law.[20] Instead, how these duties are met relies on the contracting parties or the superintendent acting in fidelity to the bargain (transparently negotiated) and orienting themselves to the individual circumstances of the other party. In other words, the demonstration of benevolence underpins contractual attempts at shaping cooperative, reasonable and honest decision-making on contracts. The original ABI model frames this as behaviour operating beyond an 'egocentric profit motive'.[21]

The concept of 'good faith' in contract law provides a similar expression of this positive orientation towards the trustor. In particular, the articulation of 'good faith' in *Macquarie International Health Clinic Pty Ltd v Sydney South West Area Health Service*[22] echoes the dimensions of the original ABI Model's conception of 'benevolence' as 'loyalty, openness, receptivity [and] availability'.[23] However, the idea of fidelity to the bargain (a common goal or a meeting of the minds as shared between contracting parties) highlights an important consideration about consensus formation. As noted above, this consensus, coming after the negotiation process, needs to be arrived at transparently.

While transparency is a value that appears to emerge from value consensus and leads to demonstrations of integrity, it also permits selfhood development and the exercise of a right to influence policy decision-making. These processes are thus antecedent to value consensus, and the visibility of these processes conveys benevolence. It is from this specific form of transparency as benevolence that genuine value articulation and consensus emerge. The results are integrous standards and the functional and technical briefs that determine ability.[24]

Future regulation of AI and ADM consequently need to be cognisant of transparency requirements related to value consensus. Transparency is recognised in the Issues Paper as an important point of consideration. However, the role of transparency mechanisms as a requirement of benevolently oriented regulatory structures has yet to be fully research, either in the analogue or the AI and ADM settings. As outlined above, our research indicates that greater understanding of transparency as benevolence is required in order to build sustainable regulatory structures that are capable of adapting to continual societal changes wrought by AI and ADM development.

2.3.2 Transparency mechanisms that build as seams into ADM process

In previous research, we framed benevolence as a positive orientation towards the individual within the dataveillance forces of automated technologies.[25] The application of benevolence as a positive orientation necessitates protecting spaces for individual development in AI and ADM systems that are built on increasingly ubiquitous data collection processes. Information privacy law produces 'seamful stopgaps' within ADM data-extraction processes that give space for individuals to 'undertake activities of self-definition and understanding'.[26] A benevolently focused regulatory structure of ADM consequently involves involving process-visibilising laws that provide more transparent frameworks of operation.[27]

One of the clear risks of AI and ADM is that technology developers—with increased access to data on the individual—can drive new personalised insights that can be used to shape individual actions and thus individuality itself.[28] In order to achieve a benevolence based regulatory approach, technology

---

19 *Augmenting Superintendent Discretion* 145.
20 Ibid 146; see, for example, the cases cited at 145–6.
21 Mayer, Davis and Schoorman, n **Error! Bookmark not defined.**, 718–9, cited in *Augmenting Superintendent Discretion* 144.
22 [2010] NSWCA 268, cited in *Augmenting Superintendent Discretion* 145.
23 *Augmenting Superintendent Discretion* 145.
24 Ibid 362.
25 *Automating Trustworthiness in Digital Twins* 360
26 Ibid.
27 Ibid.
28 Brydon Wang, 'The Seductive Smart City and the Benevolent Role of Transparency' (2021) 48 *Interaction Design and Architecture(s) Journal* 100, 106.

developers and proponents of AI and ADM must be aware of the dataveillance-to-decision processes that hides automated decision-making processes behind a black box.[29] In that regard, as highlighted throughout our submission, information privacy law provides the possibility for important and needed protections that build seams or stopgaps into the automated data flows of AI and ADM.

2.3.3 Transparency mechanisms that build on mutual vulnerability requirements

Transparency is factor of trustworthiness as benevolence that requires parties to be mutually vulnerable.[30] Vulnerability in this context includes the visibility of party-negotiating positions, even going as far as to acknowledge unequal bargaining positions. While a party may hide any misalignment in values to appear integrous (such as hiding any misaligned perceptions of the bargain struck through ambiguity in drafting), benevolence data takes on a more significant role as the contracting relationship matures.[31]

Benevolence is captured within the legal perspective of contract law and, particularly, the use of trusted intermediaries as superintendents on construction contracts. However, as these intermediaries are augmented and in certain scenarios automated across segments of the contract-administration process, benevolence considerations shift to the human-to-machine interface. In particular, benevolence requirements influence how data practices are designed and coded into these technological systems as the trust scenario scales up from private contracting to policy implementation. Accordingly, as noted above, there is an intrinsic link between the upscaling of AI and ADM that impacts upon on-ground operation and the formulation of policy that governs operation.

For example, in our research relating to the implementation of the COVIDSafe app,[32] we used the frame of mutual vulnerability to critically consider what would have been a more trustworthy deployment of the app from a benevolence-based perspective. Mutual vulnerability between the Australian Government and Australian citizens existed in three ways. First, there was uncertainty within the policy context given the nature of the global pandemic. Second, the Commonwealth's voluntary approach to limiting access to data in its contact tracing app left it vulnerable to an increased public appetite for higher regulation and decreased access to data. Third, if its policy implementation and deployment of the contact tracing app was not successful, the government would be exposed to reputational harm. Given the lethal nature of the pandemic and the assumptions the government had made in its decision-making, we contended that the deployment of the app was unsuccessful as it did not meet community expectations concerning trustworthiness.[33]

Our research revealed that the rhetorical campaigns deployed alongside the release of the COVID-19 contact-tracing app 'were not the value consensus seeking signal that benevolence characteristics require'.[34] Instead, the rhetorical campaign operated on preconceived notions that made assumptions about what the community valued. This misalignment of the federal government's two approaches (regulatory rationale and political rhetoric) blunted its ability to convey benevolence. It lacked transparency about how it was formulating a regulatory rationale to increase trust in the community in the COVIDSafe app. Accordingly, the trustworthy element of benevolence was not demonstrated because the moral compulsion diminished the perception of mutual vulnerability between the government as trustee and the urban occupant as trustor.

Our research highlighted that the sole focus on enhancing legal protections as a demonstration of value congruence was not a strong enough signal of benevolence to support trust formation. Rather, there was a need to demonstrate mutual vulnerability—not just to the ongoing health, economic and social risks posed by the pandemic but to how the government would be vulnerable to an increased appetite

29 *Automating Trustworthiness in Digital Twins* 361.
30 *Implementing COVIDSafe* 12.
31 Mayer, Davis and Schoorman, n 4, 722 (Proposition 4).
32 *Implementing COVIDSafe.*
33 Ibid.
34 Ibid 10.

for regulatory protection in information privacy protection and increased regulatory burden built on notions of fairness and proportionality.[35]

A transparent and mutually vulnerable approach would have been for the government to acknowledge these risks, thus rendering them visible to the public. By making both public and government workings more visible, there would have been a more marked demonstration of benevolence through showing a personal orientation towards the individual app user, especially in this voluntary context. This demonstration arguably would have bolstered trust and resulted in the trusting behaviour of downloading the COVIDSafe app.[36]

Our analysis of the COVIDSafe app demonstrated that regulation without clear articulation of the value consensus around data-focused technologies (including AI or ADM) will not convey necessary signals of trustworthiness. This approach then limited the opportunity to solve a clear public policy problem through the adoption of such data-focused technologies. Our analysis again highlights the need for more detailed considerations about the role of trustworthiness, and in particularly, the role of benevolence as an overarching and underpinning basis for future regulatory structures involving AI and ADM.

2.4 Are there international policy measures, legal frameworks or proposals on AI or ADM that should be considered for adoption in Australia? Is consistency or interoperability with foreign approaches desirable?

There are current attempts at making AI more explainable, fair and trustworthy – particularly the framework being adopted in the EU. This should be considered in further detail for adoption in Australia. In our view, it would be desirable to have consistency and interoperability with foreign approaches although the integrative model of trustworthiness we have set out in this submission goes one step beyond this and will position Australia as a leader in digital economy regulation.

Our submission focuses on academic developments regarding trustworthiness to better inform how the concept is being formulated in recent international work. There are competing views on what is trustworthy, and specifically, benevolent, that provide policy makers with an opportunity to examine how value consensus around machine-based expressions of benevolence is occurring. For example, Jobin et al point to over sixty guidelines globally aimed at establishing trustworthy AI (TAI) principles for developmental and deployment processes.[37] These guidelines take an interdisciplinary approach to increasing the development of TAI, focusing on making AI more explainable[38] and more suited to serving our desire for greater equality and distribution of benefits.

Thiebes et al[39] articulate five TAI principles: beneficence, non-maleficence, autonomy, justice, and explicability. They use these five principles to study seven emerging frameworks/guidelines[40] for

---

35 Ibid 12.

36 Ibid.

37 A Jobin, M Ienca and E Vayena, 'The Global Landscape of AI Ethics Guidelines' (2019) 1(9) *Nature Machine Intelligence* 389.

38 For example, see the EU Ethics Guidelines for Trustworthy AI ('EU TAI Guidelines'), OECD Principles of AI, Asilomar AI Principles, and UK AI Code.

39 Scott Thiebes, Sebastian Lins and Ali Sunyaev, 'Trustworthy Artificial Intelligence' (2021) 31 *Electronic Markets* 447, 451-5.

40 These seven frameworks / guidelines are: *Asilomar AI Principles* (Future of Life Institute, 2017) <https://futureoflife.org/ai-principles/>; *Montreal Declaration of Responsible AI* (Université de Montréal, 2017) <https://www.montrealdeclaration-responsibleai.com/>; UK Artificial Intelligence Committee, *AI in the UK: Ready, Willing and Able?* (Report, House of Lords Paper 100, 16 April 2017) ('UK AI Code'), L Floridi et al, 'AI4People—An Ethical Framework For A Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28(4) *Minds and* Machines 687 ('AI4People'); *Ethics Guidelines for Trustworthy AI* (European Commission Independent High-Level Expert Group on Artificial Intelligence, 2019) <https://digital-

achieving value consensus around how we can develop and deploy trustworthy automated decision-making processes. There is a degree of overlap between them.

- *Beneficence* describes scenarios where technology developers and powerholders deploying AI focus on enhancing human well-being, including advancing human rights and the environment in which human life occurs. Thiebes et al observe that while certain guidelines such as the UK AI Code only focus on human subjects of AI, other guidelines such as the OECD Principles on AI and the EU TAI Guidelines extend beneficence to nature and climate resilience, and even to enhancing economic life. Beneficence is related to the fields of ethical computing and AI ethics, which seek the '[promotion of] wellbeing into AI at the design and development stages'.[41]

- *Non-maleficence* describes the avoidance of harm. Thiebes et al suggest that as misuse of data leads to harm, non-maleficence is specifically oriented towards protecting information privacy, which all TAI guidelines orient towards. This principle requires technology developers to 'sincerely adhere to ethical and other pre-defined principles' and for AI systems to act 'honestly and consistently'.[42]

- *Autonomy* refers to a general promotion of 'human autonomy, agency, and oversight' where humans are given the ability to 'decide at any given time'.[43] Autonomy relates to a form of 'openness', where the system is designed to be able to 'give and receive ideas' as a means of communicating trustworthiness.[44] However, this particular principle is not unanimously captured as a key aspect within the selection of TAI guidelines studied. Instead, for the guidelines that do not accept human autonomy as a key principle, human autonomy is balanced against the other technical goals of the AI system.[45]

- *Justice* describes a spectrum of ethical goals set out within each of the TAI guidelines. Justice can take many forms, including the development and deployment of AIs to address and 'amend past iniquities like discrimination'; more equitable distribution of benefits; or the prevention of new harms. Justice considerations aim at removing bias or 'quantifying the fairness or absence thereof in AI-based systems'.[46]

- *Explicability* describes two requirements: the ability for AI to be explained and understood by human users and human subjects and the ability for AI (and the technology developers and powerholders) to be held accountable for predictions and decision outputs. However, Thiebes et al, who observed that within the TAI guidelines studied the degree of explicability varied, generally adopt the term 'transparency' for the mechanism by which TAI is realised. This principle of explicability is the 'most prevalent theme in contemporary AI research' as a reaction to the opaque nature of AI-based systems and how these systems are 'often inaccessible and non-transparent to humans'.[47]

These five TAI principles can be mapped to an extent against the trustworthiness factors they trace from the original ABI model through to its subsequent iterations aimed at automation and autonomous

---

strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>, OECD Principles on AI (OECD, 2019) <https://www.oecd.org/going-digital/ai/principles/>, *Governance Principles for the New Generation Artificial Intelligence* (Chinese National Governance Committee for the New Generation Artificial Intelligence, 2019) <https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>; and Russell T Vought, 'Guidance for Regulation of Artificial Intelligence Applications (Memorandum, M-21-06, White House, 17 November 2020) ('White House AI Principles').

41 Thiebes, Lins and Sunyaev, n 39, 451–2.

42 Ibid 454.

43 Ibid.

44 Ibid, citing AK Mishra, 'Organizational Responses to Crisis: The Role of Mutual Trust and Top Management Teams' (PhD Thesis, University of Michigan, 1992); see also PL Schindler and CC Thomas, 'The Structure of Interpersonal Trust in the Workplace' (1993) 73(2) *Psychological Reports* 563.

45 Thiebes, Lins and Sunyaev, n 39, 454.

46 Ibid, referring to Bellamy et al 2019. Thiebes, Lins and Sunyaev suggest that 'much of the current research relating to the justice principle is conducted in medical contexts' (at 455).

47 Thiebes, Lins and Sunyaev, n 39, 455.

contexts. This submission proposes that these five TAI principles could be re-organised within the taxonomy of trustworthiness factors set out in the ABI model for greater clarity. Importantly, as benevolent processes operate to generate the values that underpin demonstrations of integrity and the technical requirements to define ability, the TAI principles can also be reorganised to articulate the three themes of transparency that define benevolent design and deployment of AI and ADM. The re-organised TAI principles could consequently form a framework for the development of a trustworthiness based regulatory framework.

For example, the TAI principles of *beneficence* and *autonomy* put the human at the centre of automation technologies, imbuing the system with a personal orientation that demonstrates the trustworthiness factor of benevolence. This benevolent orientation towards the human subjected to the system leads to the articulation of social norms and standards that establish the TAI principles of *non-maleficence* and *justice*. These set out the shared values that technology developers and powerholders need to align in their design and developmental processes to demonstrate integrity. The values orient the development of these technologies towards ensuring information privacy protections remain relevant and that the system works to distribute the benefits equitably.

The TAI principle of autonomy highlights the need for seamfulness in the design and deployment of automated decision-making processes to ensure that there is sufficient space within these automated processes for human discretion, autonomy and selfhood-development. Together, these TAI principles underscore the importance of benevolent processes for the genuine achievement of value consensus.

Potentially these TAI could be codified into law and regulation, and into the program code of the technology. If so, the TAI principles of *explicability* and *beneficence* (as a form of helpfulness) provide examples of how the functions of AI and ADM systems can be more transparent, accountable, and could again form the basis for regulatory development.

There is a clear consensus emerging globally of a need to regulate the design and deployment of AI and ADM to ensure that these systems are trustworthy. In order to position Australia as a leader in digital economy regulation, we submit that the regulation of AI and ADM be enacted from a clear position of benevolence and through a specific form of transparency. Such a regulatory approach will not just regulate AI and ADM, but also define how and what regulation is enacted through value consensus, an appreciation of the crucial role information privacy laws plays in providing ‘seams’ in automated decision-making processes, and by regulators and technology developers relinquishing opaque data collection and governance practices and embracing vulnerability. In our view, this benevolent regulatory approach will enhance the likelihood of AI and ADM being designed and deployed in trustworthy ways, increasing adoption and trust in these technologies.

3. Contributing authors

This submission has been written by the following authors:

Brydon T. Wang
Lecturer, TC Beirne School of Law, University of Queensland

https://www.linkedin.com/in/brydonwang/

Brydon Wang is a lawyer and scholar researching at the confluence of technology, law and sensorised cities. He has practised as a technology and construction lawyer with top-tier law firm Allens, and recently co-edited a book, 'Automating Cities: Design, Construction, Operation and Future Impact' (Springer, 2021). His research interest lies in regulating to enhance trustworthiness in the design and deployment of automated decision-making systems in cities (BIMs, Digital Twins) and the automation of infrastructure delivery. Brydon also holds a Master of Public Policy and Management.

Dr Mark Burdon
Associate Professor, School of Law/Digital Media Research Centre, Queensland University of Technology

https://www.qut.edu.au/about/our-people/academic-profiles/m.burdon

Mark Burdon's primary research interests are privacy, information privacy law and the regulation of information security. He focuses on the complex privacy issues that arise from the sensorisation of everyday devices and infrastructures. These issues are explored significantly in his new book, *Digital Data Collection and Information Privacy Law*, published by Cambridge University Press. The relationship between privacy and power is a consistent theme in his work. Sensor data generation is instrumental to the formulation of new power relationships in networked societies. Information privacy law will consequently need to adapt. How it adapts is therefore a crucial question to resolve as we move into an increasingly collected world, where data about everything is collected and analysed. Recent research examines the information privacy law challenges underpinning the Consumer Data Right which was published in the leading data protection law journal, *International Data Privacy Law*.