Governance for Artificial Intelligence (AI) and Interoperability: Questions of Trust

Allison Wylde
Data Science for Common Good Research Group
Glasgow Caledonian University, London, UK

allison.wylde@gcu.ac.uk

Abstract: Although the rapidly emerging capabilities of AI bring potential benefits that could be transformative for cyber security, significant threats have emerged that continue to grow in impact and scale. One proposed solution to addressing important risks in AI is the emergence of strategies for AI governance. Yet, as this conceptual early-stage research argues, what is crucial for individuals, businesses, public institutions, including the military, and for high-risk environments, are questions concerning trust in AI governance. Will governance of AI be trusted? As an example, during 2023, several AI governance initiatives and strategies emerged, with some nation states proposing legislation while others looked to treaties and collaboration as solutions. Indeed, at a supra-national level, the United Nations expert multinational stakeholder Policy Network on AI (PNAI) formed to examine key issues in current AI governance. These include the interoperability of governance, data governance mechanisms, AI in supporting inclusion and the transition of nations. To help our understanding of trust in AI governance, the focus for this paper is limited in scope to interoperability in AI governance. Interoperability encompasses different aspects, policy initiatives (such as frameworks, legislation, or treaties), systems and their abilities to communicate and work together. The approach taken in this early-stage research is framed as questions of trust in AI governance. The paper therefore reviews the nature of different AI governance strategies developed and implemented by a range of key nation states and supra-national actors. This is followed by an evaluation of the role of trust, focused on AI governance strategies, in the context of interoperability in AI governance. Trust-building strategies are also considered, with a focus on leveraging the separate elements involved in trust-building to assist our understanding of the implementation of trusted AI governance. The contribution of this early-stage research is to highlight issues that may not be considered by the technical community and to contribute to developing a platform and a research approach that informs policy- learning for institutions, practitioners and academics.

Key words: Trust, United Nations Policy Network for AI, Interoperability, fit for Purpose, Policy Learning

1. Introduction

UN Secretary-General Antonio Guterres said at the January 2024 Davos meeting that AI had enormous potential for “sustainable development” but added that “every new iteration of generative AI increases the threat of serious unintended consequences” (Guterres, 2024). In seeking to address AI issues the UN formed a high-level group on AI governance (UN AI Advisory Body, 2023). However, a key problem is the lack of interoperability in AI governance among the multiple jurisdictions (UN PNAI, 2023).

The aim of this early-stage research is to examine how interoperability, as part of trusted AI governance, can be better understood. The approach taken leverages well-established trust research to allow a policy activity, reliant on trust, to be assessed.

This paper does not present a systematic review due to the contested nature of the central terms, AI, governance, interoperability, and trust. As scholars are not agreed on definitions for these terms (ESCAP, 2018, in PNAI, 2023, p.1: Hou, 2023) an abductive and interpretive approach was necessary. Thus, this paper contrasts with empirical work involving setting a hypothesis and following a deductive framework or through systematic searches, whether manual or based on machine learning (ESCAP, 2018, in PNAI, 2023, p.1: Hou, 2023).

This paper is structured as follows: after the introduction, the next part, Section (2), discusses AI governance with a focus on interoperability. This is followed by Section 3, where the processes involved in trust and trust building are covered. In Section 4, the methods are discussed with a focus on the rationale for the use of an abductive approach, this is followed by Section 5 where the expected findings are discussed. The final Section, (6), sets out the contribution of the paper along with promising directions for future work, limitations, and implications.

2. AI Governance: Interoperability

The key definitions for AI and governance are considered next with a view to presenting the definitions used in this paper.

2.1 AI and Governance

Although AI and governance are receiving increased and global attention, the terminology remains contested. The definitions themselves are problematic with some researchers viewing AI itself as lacking a coherent universally approved definition (ESCAP, 2018, in PNAI, 2023, p.1). Following the PNAI, the definition used in this paper views AI as the ability of machines and systems to acquire and apply knowledge to carry out intelligent behaviour (Ibid.)

AI governance has received attention, with several policy-level agencies calling for trust and trust building, for example, the UN Global Digital Compact (Wylde, 2023). Indeed, at the 2024 Davos meeting, the leaders of the EU and the UN called for trust rebuilding (Von der Leyen, 2024: Guterres, 2024). However, in contrast, some others suggest that AI regulation and standards may fail to increase trust and that governments must demonstrate how they are making industry accountable to the public and their legitimate concerns (Knowles and Richards, 2023).

Interoperability in the context of AI governance is loosely defined as the ability of systems and processes to communicate and work seamlessly together (UN, PNAI, 2023, p.13). To arrive at a definition of interoperability for the PNAI multistakeholder team, literature on interoperability and policy that was operational during the period July to October 2023 was reviewed, the definition as finally agreed is set out in Table 1 (UN, PNAI, 2023).

Table 1: Definition in AI Governance interoperability, including the factors involved (processes, activities and communications and cooperation, UN, PNAI, 2023), expanded to illustrate the primary trust referents and roles.

Definition; three interlinked factors (UN, PNAI, 2023)	Factors involved	Primary trust-referent (trust-level); and key role for trust
Processes	Tools, measures, and mechanisms	Policy-level trust; trust in policy
Activities	Multi-stakeholders and their interconnections	Organization, institution, and individual-level trust; trust in partners
Communications and cooperation	Agreed ways (Multi-stakeholders and their interconnections)	Organization, institution, and individual-level trust; trust in communication and cooperation

Three key interlinked factors were identified as being important in specifying a definition for interoperability in AI governance (UN, PNAI, 2023). Table 1 summarises the definition of interoperability, agreed as follows, interlinked factors, including: (i) processes, comprising tools, measures, and mechanisms (ii) activities undertaken by multi-stakeholders and their interconnections and (iii) agreed ways to communicate and cooperate (UN, PNAI, 2023, p. 13).

3. Trust

What follows is not a systematic review. For the purposes of this paper, which is focused on questions of AI governance, prominent trust theory, from organization and management literature, the integrative trust model (ITM) (Mayer et al., 1995), is leveraged. This approach provides a conceptual framework through which issues of interoperability can be understood. As Table 1 illustrates, each of the three factors involved in interoperability are viewed as founded on trust, even though trust and the primary trust referent are not specified.

Trust is a well-researched construct with studies across multiple disciplines. Due to the sheer volume of research on trust, this paper is limited to the view of trust as a relational and subjective phenomenon (Mayer et al., 1995: Rousseau et al, 1998). Indeed, as Mayer et al. (1995) find, the referents of trust are often not specified. Thus, evaluation of relational-trust is well-suited to the application of theory drawn from management and organization studies (Hou, 2023).

In this perspective, relational trust is seen as an individual taking a decision to trust, based on antecedents, an assessment of trust and trust-building (Mayer et al., 1995). At the level of inter-person trust formation, a trustor is viewed as holding positive expectations that a trustee will perform an action valuable to the trustee, irrespective of control (Mayer et al., 1995). The assessment of the trustee is conducted on the characteristics of

ability, benevolence and integrity moderated by the trustor's propensity to trust. In the final stage, the trustor accepts vulnerability and takes a risk in placing trust in the trustee (Rousseau et al., 1998).

Trust occurs across different levels, from inter-person trust to trust in teams, trust in organizations and trust in institutions (Fulmer and Gelfand, 2012). Trust also occurs beyond trustor to trustee relations as trust in technology (McKnight, 2011) or trusted AI (Wylde, 2022a) and trust in robots (Hou, 2023). Researchers have also specified a role for trust in institutional policies, even though a trustee may have little experience an institution, for example, a government (Möllering, 2013). Institutional level trust has been shown to be supported by processes such as tax or legal systems and the trusted individuals involved in these processes (Vanneste, 2016) that serve to reduce vulnerability and uncertainty (Rousseau et al., 1998: Möllering, 2023).

The definition of trust for this work involves antecedents; that a trustor possesses positive expectations, then a process of trust assessment, based on ability, benevolence and integrity followed by outputs involving trust-building (Mayer et al., 1995). It is important to note that this definition includes an acceptance of vulnerability as part of trusting (Rousseau et al., 1998). As a final comment, the process of trusting is moderated by the trustor's individual propensity to trust (Mayer et al., 1995).

Building on this definition, a conceptual framework is proposed that can examine trust as a construct in processes such as interoperability, involving trust among a range of interacting actors, policy, and processes (Oomsells and Boukaer, 2014) set out in Table 1. This framework is next applied to separate out the complex processes involved in examining trust in a process such as interoperability.

Research into trust in AI globally has identified 61% of people reporting not trusting AI and 71% expecting AI to be regulated, but only 33% lacking confidence in governments and businesses to develop, use and regulate AI (Gillespie et al., 2023). However, why this may be the case has not been explored in depth (Knowles and Richards, 2023). Further research has identified that stakeholders are contacted differently: those most familiar with technology are more likely to be assessed sooner and they are found to be the most comfortable with AI, while conversely, disadvantaged and vulnerable groups are less likely to be contacted and more likely to be the least comfortable with the technology (Knowles and Richards, 2023).

Trust in AI is considered as layered, with trust involved in several domains: data, the technology and platforms, the supervisors and the users, developers and the organizations that deploy the AI, the regulators and an important consideration is the domain of the application (Knowles and Richards, 2023).

Summing up, as clear gaps for qualitative research into the who, why and what of trust in AI have been identified, the call for further research is picked up in this paper.

4. Research Method, Analysis and Preliminary Findings

A research approach involving interpretation was followed to allow the interlinked concepts to be teased-out (Hou, 2023). The research approach is discussed followed by a summary of the first preliminary findings.

4.1 Analysis

The rationale for this approach is due to the nature of the research question and the study material. Following well-established conceptual practices, the study adopted an abductive approach. For this type of study, which relies on interpretation and making meaning, a deductive and central tendency approach is not appropriate due to the lack of agreement among scholars on issues such as definitions. In consequence, research approaches involving hypothesis testing or a machine learning driven systematic searches are not readily supported (Hou, 2023). In addition, as this is early-stage research, the scope of is limited to the UN PNAI and High-level AI (PNAI, 2023: UN AI Advisory Body, 2023).

Prominent management researchers and editors recommend approaching problems of meaning using through theory building or extension by abductive approaches since "we don't know what we think until we know what we write" (Forster 1927, in Byron and Thatcher, 2016). Following an iterative abductive process, the study progressed as follows. First the author developed a foundational idea to tackle the question of creating understanding in the issue of AI governance, in particular, that of interoperability (Sætre and Van De Ven, 2021). The central idea in this paper is that interoperability in governance can be framed as questions of trust (Mayer et al., 1995). This idea is justified through recognizing the ability of the trust literature, in particular, the integrative trust model (Mayer et al., 1995) in offering a model through which a phenomenon such as governance (policy) could be examined (Sætre and Van De Ven, 2021). Abduction is suited to questions

concerning understanding our world (King and Kay, 2020) as it "becomes more dynamic, interconnected and uncertain" (Sætre and Van De Ven, 2021, p.684),

Next considered are the processes involved in interrogating the theory. The author followed Byron and Thatcher (2016); visual representation, tables, charts, and notes were created to help tease out the key elements involved in building the theoretical framework and examining the important elements and processes. The social processes then included informal presentations of draft work for feedback followed by discussions with colleagues and iterative development of the work. Some avenues that appeared promising at the start of the process were followed further others were altered to refocus on the salient research questions. In this way success and failure in research were considered and integrated into the research process (Sætre and Van De Ven, 2021). The overarching aim as central to abductive research was to create plausible and meaningful material that could form findings (Sætre and Van De Ven, 2021).

4.2 Analysis and Materials

The rationale for the selection of the analysis approach and selection of study materials is based on the limited scope of this early-stage research. Although methods for trust research includes a comprehensive range of approaches for trust studies, including trust scales (Gillespie, 2011). As this early-stage research study is restricted to evaluation involving the assessment part of the ITM model (Mayer et al., 1995). Again, it is acknowledged that the scope of the sample analysed does not reflect the entire landscape, and it is acknowledged that some regions disproportionally contribute to the development of policy (PNAI, 2023). As outlined in the scope, for this early-stage work, the review is limited to the UN documents produced by the UN PNAI and the UN AI expert group (PNAI, 2023: UN AI Advisory Body, 2023).

4.3 Preliminary Findings

Key themes emerging from the study thus concern the institutional stakeholders involved, the actions to be undertaken (norms, rules, standards) and processes and mechanisms to be implemented. In terms of trust, the issues concern consistency in the terms, for building or promoting trust or in addressing declining trust.

Table 2: Preliminary findings: examples of the research material, key themes, trust level and referents (UN AI Advisory Body, 2023).

Examples from research material (UN AI Advisory Body, 2023).	**Key themes**	**Trust level/ referents**
The UN Advisory Body is uniquely placed to help through "turning a patchwork of evolving initiatives into **a coherent interoperable whole, grounded in universal values** agreed by its member states, adaptable across contexts" (p. 6).	The need for international cooperation to tackle AI governance	Trust in a regulator (the UN)
Recognising; no alignment either in terms of interoperability between jurisdictions or in incentives for compliance. With policy ranging from binding rules to nudges that are non-binding (p.13/14).	Lack of policy alignment among different jurisdictions	Trust in a regulator (the UN), trust in policy
A simplified schema is presented for considering the emerging AI landscape, which the Advisory Body say they will develop further (p. 13).	Need for terminology	Trust in a regulator (the UN), trust in policy
Awareness amongst existing states and the private sector, call for new organization structure to be entrusted (p. 16).	Need to create a new organization	Trust in a regulator (the UN), trust in policy
Grounding in norms Actions to reinforce interoperability include grounding in international norms in a universal setting (p.18).	Agreed policy type (norms)	Trust in a regulator (the UN), trust in policy
Fora could include the UN organizations and For a such as UNESCO and ITU to reinforce interoperability; global membership the UN can bring states together, develop common socio-technical standards, ensure legal and technical	Driver organizations	Trust in a regulator (the UN), trust in policy

Examples from research material (UN AI Advisory Body, 2023).	**Key themes**	**Trust level/ referents**
interoperability; balance technical interoperability with norms (19).		
Actions involving "Surfacing best practices for **norms and rules**, including, for risk mitigation and economic growth. Align, leverage, and include, soft and hard law, standards, methods, and frameworks developed at the regional, national, and industry level to support interoperability" (p.23 12-24months).	Agreed policy type (norms)	Trust in a regulator (the UN), trust in policy
Ensure interoperable action at all levels – **across all institutions, frameworks** (national and regional) and the private sector (p.24).	Agreed policy type	Trust in a regulator (the UN), trust in policy
The UN will **pursue research** on risk assessment methodologies and governance interoperability (p.25).	Need for research	Trust in policy

Further detailed examination of the research material will be undertaken to identify themes as they may align with trust (or not) and consider the interconnections across themes. The aim is to create findings that help highlight important directions that can help policy makers as they develop AI governance policy.

5. Conclusions

The contributions of this early-stage paper are twofold. Firstly, the gap in our understanding of AI governance from the perspective of a lack of interoperability is addressed through identifying the need for institutions to demonstrate they hold the private sector accountable and that they acknowledge their stakeholder's concerns, focused on vulnerable stakeholders (Von der Leyen, 2024). Secondly, a method has been proposed to handle the contested nature of the central terms and the lack of consistency. The framework is based on an interpretive abductive approach applied to build understanding and leverage trust theory to understand operationalizing trust viewed as helping to achieve interoperability in AI governance.

As with all research, limitations are present. In the trust theory presented, trust is viewed as a linear input-output process, starting from antecedents to the assessment of trust, and finally to trust-building (Wylde, 2022b). As such a simplistic process fails to account for the dynamic and simultaneous nature of trust encounters (Dietz, 2011). This limitation could be taken up in future research that could unravel the nature of the interlinked-processes and sequences involved in trust decision-making. Future research could use machine learning to review text calibrated through multiple perspective or management and organization studies trust theory, helping to refine the constructs. Additional attention could focus on terms such as trust building and addressing trust deficits (Von der Leyen, 2024).

It is hoped that this early-stage work provides a foundation that can be built upon to help policy makers as they grapple with the complexities involved understanding and achieving trusted AI governance, in particular, issues of interoperability. As ever, a call goes out for further research on trust and interoperability in this increasingly important and contested domain of AI governance.

References

Byron, K. and Thatcher, S.M. (2016) Editors' comments: "What I know now that I wish I knew then"—Teaching theory and theory building, Academy of Management Review, 41(1), pp.1-8.

Dietz, G. (2011) "Going back to the source: Why do people trust each other?" Journal of Trust Research, 1 (2): 215-222.

ESCAP, UN. (2018) "Enhancing cybersecurity for industry 4.0 in Asia and the Pacific", [online]. https://repository.unescap.org/handle/20.500.12870/238 [Accessed, 24. Jan. 2024].

Forster, E.M. (1927) Aspects of the Novel. Harcourt, Brace.

Gillespie, N. (2011) "Measuring trust in organizational contexts: an overview of survey-based measures", Handbook of research methods on trust, p.175.

Gillespie, N., Lockey, S., Curtis, C., Pool, J. and Akbari, A. (2023) "Trust in Artificial Intelligence: A global study", The University of Queensland and KPMG Australia, doi: 10.14264/00d3c94

Guterres, A. (2024) At Davos forum, Secretary-General warns of global norms collapsing, highlights the need to rebuild trust reform governance, [online]. https://press.un.org/en/2024/sgsm22109.doc.htm [Accessed, 24. Jan. 2024].

Hou, M. (2023) “Challenges in Understanding Trust and Trust Modelling : Quenching the Thirst for AI Trust Management”, In Transactions on Computational Science, XL, pp. 1-5. Berlin, Heidelberg: Springer Berlin Heidelberg.

Knowles and Richards. (2023) “Trusted AI”, Association for Computing Machinery (ACM), Technology Policy Council, TechBriefs, [online]. https://dl.acm.org/doi/pdf/10.1145/3641524 [Accessed, 24. Jan. 2024].

Lewicki, R.J., McAllister, D.J. and Bies, R.J. (1998) “Trust and distrust: New relationships and realities”, Academy of Management Review, 23(3), pp.438-458.

Mayer, R., Davis, J. and Schoorman, F. (1995) "An integrative model of organizational trust", Academy of Management Review, 20(3), pp. 709-734.

McKnight, D.H., Carter, M., Thatcher, J.B. and Clay, P. (2011) "Trust in a specific technology: an investigation of its components and measures", ACM Transactions on management information systems, 2(2), pp. 1-25.

Möllering, G. (2013) “Trust without knowledge?” Comment on Hardin, ‘Government without trust’, Journal of Trust Research, 1(1), pp. 53-58.

Oomsels, P. and Bouckaert, G. (2014) “Studying interorganizational trust in public administration: A conceptual and analytical framework for" administrational trust", Public Performance and Management Review, 37(4), pp. 577-604.

Rousseau, D.M., Sitkin, S.B., Curt, R.S. and Camerer, C. (1998) "Not so different at all: a cross discipline view of trust", Academy of Management Review, 23(3), pp. 393-404.

Sætre, A.S. and Van de Ven, A. (2021) “Generating theory by abduction”, Academy of Management Review, 46(4), pp.684-701.

UN AI Advisory Body. (2023) “Governing AI for Humanity”, Interim Report. Dec. 2023, [online].
https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf [Accessed, 24. Jan. 2024].

UN Policy Network on Artificial Intelligence (PNAI). (2023) “Strengthening multistakeholder approach to global AI governance, protecting the environment and human rights in the era of generative AI”, in Sipinen, M. (Ed.) United Nations Internet Governance Forum, [online]. https://intgovforum.org/en/content/pnai-work-plan [Accessed, 24. Jan. 2024].

Von der Leyen, U. (2023) “Special address by President von der Leyen at the World Economic Forum”, 16. Jan. 2024, [online]. https://ec.europa.eu/commission/presscorner/detail/en/speech_24_221 [Accessed, 24. Jan. 2024].

Vanneste, B.S. (2016) "From interpersonal to interorganizational trust: the role of reciprocity", Journal of Trust Research, 6(1), pp. 7-36.

Wylde, A. (2022a) “Cyber Security Norms: Trust and Cooperation”, Conference paper. ECCWS 2022.

Wylde, A. (2022b) “Questions of trust in norms of zero trust”, In Intelligent Computing, Proceedings of the 2022 Computing Conference, 3, pp. 837-846. Cham: Springer International Publishing.