WEP Warwick Evans Publishing

International Journal of Social Sciences and Public Administration

ISSN: 3005-9836 (Print), ISSN: 3005-9585 (Online)
Volume 10, Number 9, Year 2026
DOI: https://doi.org/10.62051/ijsspa.v10n9.06



Analysis of the Construction of Artificial Intelligence Trust and the Practice of AI Governance under the Background of Digital Transformation

Yuhan Du

Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia

ABSTRACT

The rapid diffusion of artificial intelligence in many domains has transformed the problem of trust from the psychological acceptance of technology by the individual to an institutional difficulty throughout the process of technological socialisation. This study covers the progress of China's AI governance policies from top-level design to systematic regulation, concentrating on the governance transformation represented by the "Artificial Intelligence Security Governance Framework" 2.0 version and the modification of the "Cybersecurity Law". The analysis shows that the trust challenge is mostly caused by three kinds of contradictions, including the poor explainability of technology, the possible concerns of data privacy security, and disagreement on the fairness of algorithms. One of the primary trends in governance practice today is the movement from static management of risk classification to dynamic reactions to evolving risks, with security considerations gradually established in the technology lifecycle. The fundamental principle of this transformation is to respond to the uncertainty of technological innovation with institutional flexibility. This study claims that to establish confidence in artificial intelligence we need to concentrate on institutional credibility and technical reliability, and the key is balancing innovation incentives and risk prevention.

KEYWORDS

Artificial Intelligence Trust; AI Governance; Security Governance Framework; Trust Construction; Dynamic Governance.

1. INTRODUCTION

Artificial intelligence has grown from an adjunct to a decision-making subject. The 2025 Government Work Report proposed to continue to promote the "artificial intelligence +" activity and promote the large model to be widely deployed. Technology is applied from industry to social governance. Intelligence is always developing in sectors like government affairs, finance and health care. But the progress of technology has not automatically resulted in greater social trust. The public sees increased efficiency but also mistrust of trustworthiness because of the credibility of generative artificial intelligence, the opaqueness of algorithms and data security risks.

Trust is an obstacle to the social usage of technology [1]. Algorithms used in banking, healthcare and governance are raising trust issues due to fairness, the inability to explain diagnoses and a lack of openness about the decisions being made. Confidence in artificial intelligence is a question of the allocation of technological power, of attribution of responsibility, of the foundation of institutional legitimacy.

Current research is mainly technology-driven, aiming at solutions that improve explainability and robustness to generate confidence. Such methods restrict the problem to technology answers, while

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

disregarding the social construction aspects of trust. When the general people cannot understand the results of algorithms and begins to doubt, this exposes the fact that there is an asymmetry of cognitive power between creators and users. The trust issue needs to be embedded into the AI governance institutional frame to address this issue.

In this study, we investigate the subject of how governance approaches might lead to trust in artificial intelligence. What is the institutional evolution of China's AI governance system? What has been done regarding the present governance arrangements? What are their shortcomings? The paper discusses the institutional course of China's AI governance policies from top-level design to specific implementation, analyses the structural dilemma of trust-building, investigates the paradigm shift of governance led by the 2.0 version of the "Artificial Intelligence Security Governance Framework", and deliberates the practical ways of trust-building from the industry perspective.

2. INSTITUTIONAL TRAJECTORY: THE EVOLUTION LOGIC OF CHINA'S AI GOVERNANCE POLICIES

The first step to understand the growth of confidence in artificial intelligence is to look at the formation of governance structures. The institutionalisation of AI governance in China has been a journey from concept to rule, and from dispersion to integration. In 2017, "New Generation Artificial Intelligence Development Plan" first included law and ethics into the institutional system of artificial intelligence development [2]. This paper did not provide much tangible legislation but it did validate a principle at the level of the national strategy: the evolution of artificial intelligence cannot be left to the market and technology to progress alone, the system must evolve in concert as well. In 2021, the "Plan for the Construction of the Rule of Law in China (2020-2025)" also proposed to strengthen laws in the sphere of information technology, and put the construction of legal systems in linked businesses on the agenda. The judgement of these two works is that governance is not for the restraint of technology, but rather a necessary precondition for the healthy development of technology.

The high level design sets the direction, while specialised legislation governs implementation. In particular, laws witnessed China taking a strategy of advancement in the many fields. In terms of algorithm governance, the "Regulations on the Management of Algorithmic Recommendation for Internet Information Services" in 2022 and the "Regulations on the Management of Deep Synthesis for Internet Information Services" in 2023 require traceability of algorithm services, which is equivalent to pulling algorithms out of pure technical tools into the domain of accountable social behaviour [3]. The "Interim Measures for the Management of Generative Artificial Intelligence Services" (generative artificial intelligence) issued in 2023 adopted the principle of "inclusive and prudent, classified and graded", and made it clear that service providers should have media attributes or social mobilisation capabilities to carry out safety assessments and algorithm filings. In the data governance layer, the "Data Security Law" and "Personal Information Protection Law" were passed successively in 2021, setting the compliance bottom-line. In 2022, the "East-to-West Computing Resource Allocation Project" was begun, and in 2023, the "Action Plan for High-Quality Development of Computing Infrastructure" was released. The computing power resources have been driven to the national strategic level in terms of computing power infrastructure.

This institutional arrangement is characterised by "separate division and governance", with the advantage of flexibility and the ability to quickly form regulatory coverage, but the problem is also obvious: the links between each link are not tight, and there is a lack of a high-level specialised law, and the overall system is lacking in comprehensiveness. With the rising iteration of the technology of artificial intelligence (AI), this gap becomes more and more visible. Lays a framework for the upgrading of the "Artificial Intelligence Security Governance Framework" and the revision of the "Cybersecurity Law" by adding AI provisions later.

3. THE TECHNOLOGICAL ROOTS AND SOCIAL NATURE OF THE TRUST CRISIS

The first step to recognising the need for governance is understanding the disappearance of confidence. The loss of trust in AI is not a technological problem, but a structural gap, caused by the reengineering of social interactions by technology [4]. The introduction of AI between the doctor and the patient, between the court and the litigant, between the platform and the user disrupts the initial trust link. Such a change can be explained at three levels: the nature of technology, the flaws in the protection of rights and the inequality in expectations of justice.

Large language models make judgements in a completely different way than normal software does. The latter is fixable with code review. The former needs massive neural networks with trillions of parameters, very nonlinear and unpredictable. Even the developers may not be able to explain all of the reasons for each output. Explainability is vital to procedural legitimacy, not only an efficiency issue, in areas like healthcare and justice. When doctors deploy AI for diagnostic help, the doctor is effectively promising a "unclear" system, if the system can't explain why it made a conclusion. But the secrecy of the system can also cast doubt on the legitimacy of the judgement, while judges rely to the sentencing system to make verdicts. The major difficulty for faith in AI is precisely how it erodes the institutional underpinning of the "duty to explain" for professional decision-making.

The other logic is the privacy problem. The more sophisticated the AI, the more granular data it requires. The Personal Information Protection Law has set up the standard of "notification-consent", but in actuality, due to the lack of information equality, people are prone to click "agree". What is even more difficult is that the data demand of AI is not just during the training stage, but also during the user involvement with data for continual optimisation, and the privacy risk persists throughout. Users don't know where the data goes, if they can delete it. Trust is a rational choice.

The problem of algorithmic fairness reveals the biases that hide under the surface of "objectivity". Algorithm learning is the data and the biases within it. In 2025, a California court certified a class action against the Workday employment algorithm for its disproportionate impact on job candidates over 40 years of age. Also in the same year, Earnest paid $2.5 million damages for the AI underwriting model's disparate treatment of persons of colour. Human biases are more forgiving than algorithmic biases because human biases are based on individual subjective errors. But once the bias is integrated in the algorithm, the unfairness is ingrained in the technical system and may be scaled up massively [5]. Identifying the root causes of these issues could enable targeted governance.

4. TRANSFORMATION OF GOVERNANCE PARADIGM: FROM STATIC REGULATION TO DYNAMIC GOVERNANCE

The triple conundrum of technical black box, privacy difficulties and algorithmic biases has prompted a shift in governance thinking. The "Artificial Intelligence Security Governance Framework" has undergone a significant evolution from version 1.0 in 2024 to version 2.0 in 2025, transitioning from static regulation to dynamic governance. Grasping this transformation is vital to grasp the process of institutional building of trust in artificial intelligence [6].

The most intuitive is the change in risk category. In version 1.0 security issues were classified into two categories, "endogenous" and "application", depending on where they originated. In the second edition it was improved by classifying the dangers into three types: technical endogenous risks, technical application risks and risks arising from the application. Particular attention should be made to the new "application derived risks". It is not the problems of the technology itself. It is not the same as the obstacles of certain applications in some cases. And it is the structural reverberations of the widespread use of artificial intelligence: changes in patterns of labour, disturbances to the information environment and perils to cognitive security. The effect of this change is evident. The governance

perspective is extended from the two-dimensional framework of "technology-application" to the three-dimensional framework of "technology-application-society [7].

Meanwhile, a grading procedure has been set up. No apparent grading system existed in version 1.0 but version 2.0 would classify the threat levels in three dimensions: application scenarios, intelligence level and application scale. First, it gives the grading factors and classification criteria. This strategy is for differentiated governance, where high-risk scenarios are subject to strict security audits and penetration-based supervision, and low-risk scenarios are subject to simplified compliance requirements. Grading is intended to direct governance resources to the highest risk areas, and to avoid a one-size-fits-all approach that could stifle innovation.

Version 2.0 does not remain at the level of general principle guidance for governance measures but progresses to more operational institutional arrangements. The legislative route is clear: on one hand, to encourage high-level legislation, in contrast, to persuade local governments to formulate precise rules and provide criteria for the construction of technical ethics guidelines. Regulatory model Inspection to embedding through the complete process at one point The "Framework" 2.0 explicitly provides a dynamic adjustment mechanism for governance measures, and builds the governance framework itself as an open system that constantly corrects itself according to technology advancement and risk changes. The core premise is that artificial intelligence challenges are not things we find and put into a box. They are in a symbiotic connection with technology iteration.

5. FULL LIFE CYCLE GOVERNANCE: THE INSTITUTIONAL LOGIC OF EMBEDDED TECHNOLOGY

In this chapter, the detailed logic of embedding full life cycle is explicitly described. Full lifecycle governance addresses the operational issue of how to govern. This necessitates that security considerations be integrated in all phases of system design, development, testing, deployment, and operation & maintenance, thus making security an inherent constraint in the development process, rather than an external compliance requirement.

This is a response to the failure of post-event supervision in AI settings: full lifecycle governance. The traditional supervision has a sequence of "occurrence of harm-determination of liability-execution of penalty" and its success hinges on the traceability of harm and clarity of the accountable party. These two conditions are not met by the way artificial intelligence systems make judgements. This kind of fluctuation in the training data might result in systematic bias in the model output; any combination in the parameter space can result in unexpected detrimental content. When the damage is obvious, it is technologically difficult to trace the causal chain and the cost of accountability is heavy. Security limits with governance focus moving forward in design and development phases.

Another purpose of this governance technique is to translate legal principles into implementable technical needs. Article 7 of the "Interim Measures for the Administration of Generative Artificial Intelligence Services" provides that the sources of training data shall be legal, and the data shall be true, accurate, objective and diverse. This clause is a principle rule. The realisation of this provision has to set specified review nodes and compliance standards in data collection, cleaning, annotation, training and testing. The legal principle cannot be executed in development practice since there are no embedded operational nodes [8].

Full lifecycle governance is not necessarily more limited security-wise. More layers of protection means less danger, but also more development cost and slower innovation. The governance difficulty is to discover the fair limit of embedding. The Framework 2.0 version adopts the hierarchical thinking, which strikes a balance between high-risk scenarios such as government affairs disclosure, judicial evidence collection and intelligent driving requiring strict full lifecycle security audits and low-risk applications that can simplify compliance procedures. The tiering is designed to be a suitable trade-off between security and efficiency [9].

The technical infrastructure follow up also supports this option. Domestic research institutions have established the field of algorithm security assessment, and have formed assessment capabilities including data security assessment, model development audit and operation risk monitoring, and can implement differentiated security assessment plans for systems with different risk levels.

6. IMPLEMENTATION OF TRUST: INTEGRATION OF INDUSTRY PRACTICES AND INSTITUTIONAL FRAMEWORKS

The institutional frame and the technical infrastructure have laid down the general principles of governance. However, the actual success of these rules will be determined by the changes and applications that various industries will make. The risk structures and trust foundations are distinct in the three domains of banking, healthcare and government affairs, and the actual forms of AI governance are varied, and rightly so.

The banking industry was the first to start and has rather sophisticated risk management practices. In December 2025, 26 institutions jointly released the first domestic group standard for large models in the financial industry, "Reference Framework for Trustworthy Applications in the Financial Domain", which directly solves the problem of how to build, manage and evaluate large models in the financial industry when implemented[10]. The significance of this book is in the translation of the ideas of the "Framework" 2.0 edition to the relevant evaluation elements and governance techniques.

The logic of trust is different in the health sector. However, the obstacles of AI-supported diagnosis are not technical soundness, but the compatibility of explainability and doctor-patient contact. Patients have a right to know the reason for the diagnosis and professionals have an obligation to explain [11]. If the AI system can't explain the judgement thinking, doctors will be in a dilemma: use AI but can't explain, which is a violation of the obligation to inform; give up AI and may miss better alternatives. This was addressed in the "Expert Consensus on Ethical Governance of Generative Medical Artificial Intelligence Clinical Applications" published in 2025, which proposed 29 recommendations centred on four key governance risks, technical adaptation and fairness, full life-cycle supervision, and preservation of doctor-patient relationships. "The objective is to keep AI in a tool role and leave the explainability up to the doctors".

The issue of administrative legality is the state of the government affairs. Citizens have the right to know the basis on which administrative decisions are reached, and this concept does not change due to technology intervention. Hangzhou Yuhang District released the first "AI + Government Affairs" management system of Zhejiang Province in December 2025. Guangdong has led the way in conducting rigorous risk testing on "AI government service assistants" covering 12 livelihood-related categories such as social security and healthcare. Based on the behaviours of two regions, we can generally conclude that the transparency standards of government affairs scenarios are higher than those of commercial scenarios.

Capacity building for governance is also being pursued at the level of enterprises. In 2022, Alibaba set up a Technology Ethics Committee and by 2025, all key divisions of the firm had established governance committees. SenseTime has designed a review system for ethical risk classification across the full product lifecycle and produced the SenseTrust platform for model security testing. These tactics are generally used to translate external demand for compliance into internal organisational capabilities and embed governance in the quality system rather than merely reacting to inspections.

7. CONCLUSION

Artificial intelligence is about institutional provisioning, not technological advancement, and it is about trust. Explainable algorithms can help to overcome the problem of cognitive transparency, but they cannot replace the functions of institutions in terms of responsibility allocation and expected

stability. The research on institutional change indicates that the main difficulty of trust construction is not in the model layer, but in the rule layer.

The governance framework development is founded on a clear risk-awareness strategy. The level of institutional completeness has been continuously improved, from the creation of the ethical legal orientation in the development plan in 2017, to the filling in of the gaps with sector-specific regulations, the systematisation of risk classification and full life-cycle governance in the 2.0 version of the framework in 2025, to the revision of the Cybersecurity Law in 2026, which elevates AI governance to a legal issue. The major motive for the process is the constant strengthening of awareness of the nature of the threats of AI at the governance layer: first technical security issues, then the social impact of applications, and eventually the general resilience of the technology-society system.

A longer chain of institutions does not necessarily alleviate the problem of trust. The pathway from rule texts to societal trust is three steps: industry adaptation, internalisation of firms and public cognition. The financial industry is about the transformation of standards, the healthcare business is about the distribution of the right to interpret, and the government affairs sector is about the transparency requirements. These three directions indicate the need for a particular industry adaptation space in a single frame. Also, in the ethical governance practices of organisations, it is noted that external compliance requirements only get actual binding force when they are incorporated into internal processes.

Three additional sets of structural contradictions constraining the generation of trust are now the time lag between technological iteration and institutional revision; the disconnect between the global nature of the technology supply chain and the nation-state scope of the governance framework; and the disconnect between the professionalisation of governance discourse and the everyday nature of public cognition. In response to these paradoxes, it is necessary to go forward on four simultaneous dimensions: institutional flexibility, technical open source, international involvement and public education. Trust in AI is an ongoing process of adjustment and there needs to be a balance between the rigidity of norms and the flexibility of governance.

REFERENCES

[1] Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, *1*(4), 1–45.

[2] Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society*, *36*(1), 59–77.

[3] Yang, F., & Yao, Y. (2022). A new regulatory framework for algorithm-powered recommendation services in China. *Nature Machine Intelligence*, *4*(10), 802–803.

[4] Durán, J. M., & Pozzi, G. (2025). Trust and trustworthiness in AI. *Philosophy & Technology*, *38*(1), Article 16.

[5] Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic literature review on bias mitigation in generative AI. *AI and Ethics*, *5*(5), 4789–4841.

[6] Xiao, B. (2025). Agile and iterative governance: China's regulatory response to AI. *Computer Law & Security Review*, *58*, Article 106183.

[7] Currie, W. L., Leimeister, J. M., Schlagwein, D., & Willcocks, L. (2025). Rethinking technology regulation in the age of AI risks. *Journal of Information Technology*, *40*(3), 236–245.

[8] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys*, *56*(7), 1–35.

[9] Ghosh, A., Saini, A., & Barad, H. (2025). Artificial intelligence in governance: Recent trends, risks, challenges, innovative frameworks and future directions. *AI & Society*, *40*(7), 5685–5707.

[10] Adari, V. K., & Sarabu, V. B. (2021). Building trust in AI-first banking: Ethical models, explainability, and responsible governance. *International Journal of Research and Applied Innovations*, *4*(2), 4913–4920.

[11] Khan, M. M., Shah, N., Shaikh, N., Thabet, A., & Belkhair, S. (2025). Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, *195*, Article 105780.