Data & Policy (2026), 8: e48
doi:10.1017/dap.2026.10099

RESEARCH ARTICLE

An interpretable and adaptive AI governance framework

Ya Wang

Shanghai Jiao Tong University, China
Email: wangya0315@sjtu.edu.cn

Received: 22 June 2025 UTC; **Revised:** 06 May 2026 UTC; **Accepted:** 01 August 2026 UTC

Keywords: adaptability; AI governance; AI innovation; artificial intelligence; dynamics

Abbreviation: AI, artificial intelligence; DP, differential privacy; EU, European Union; IBM, International Business Machines Corporation; OECD, The Organization for Economic Co-operation and Development

Abstract

In recent years, the rapid development of artificial intelligence has raised a series of ethical issues. How to balance AI innovation and AI governance has attracted wide attention from society. Some principle-based and practical approaches have already emerged. However, these strategies mainly face problems of time, speed, and cost. They fail to fundamentally balance dynamic factors and the need for explainability. This study proposes an explainable and adaptive AI governance approach from the perspective of time to balance AI governance and AI innovation. Causal models have causal explanation capabilities. They meet the requirements of responsible innovation. They can improve governance speed and reduce governance and innovation costs. Taking Bayesian networks as an example, they have explainability and dynamic adaptation ability. Bayesian networks can be used as a governance step before a regulatory sandbox. Bayesian networks can implement a human-assisted regulatory approach. This approach can combine the EU Artificial Intelligence Act with scholars' recommendations on quantifiable risk levels. The explainable and adaptive AI governance approach has three important meanings for AI governance. First, it can provide an inclusive governance path. Second, it can supplement static governance strategies. Third, it can automatically generate explainable value alignment schemes.

Policy Significance Statement

The implications of the study for policymakers are as follows:

1. On the one hand, companies that follow this framework can shorten the time for technology release and application, which helps to avoid profit loss caused by long review periods. On the other hand, governments can use this framework to address the potential delay and static nature of traditional governance schemes.
2. Using Bayesian networks for risk assessment allows regulators to make targeted interventions and improve regulatory efficiency. At the same time, applying Bayesian networks to AI governance together with regulatory sandboxes and full-cycle human oversight provides triple protection for AI governance.
3. It assists in supporting responsible prediction and decision-making processes.
4. The explainable and adaptive AI governance approach has three important meanings for AI governance. First, it can provide an inclusive governance path. Second, it can supplement static governance strategies. Third, it can automatically generate explainable value alignment schemes.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

1. Introduction

At present, artificial intelligence is widely applied in important areas such as healthcare, manufacturing, education, and finance. It has become a key factor in improving national strength and driving economic development. At the same time, it has also raised many discussions on topics such as transparency and trustworthiness, safety and reliability, bias, fairness, responsibility, and privacy. In light of this, it is important to maintain a balance between AI innovation and AI governance.

In recent years, the European Union, the United States, China, Germany, and Singapore have introduced policies on AI governance and made some progress. China has clearly stated the need to balance AI governance and innovation. In July 2023, China's Cyberspace Administration, National Development and Reform Commission, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration issued the Interim Measures for the Administration of Generative AI Services (https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm). It clearly stated that "the country adheres to the principle of giving equal importance to development and security and combining the promotion of innovation with governance according to law. Effective measures will be taken to encourage the innovative development of generative AI. Generative AI services will be supervised in an inclusive, prudent, and classified manner."

In December 2024, the China Academy of Information and Communications Technology published the Blue Book of AI Governance (https://www.caict.ac.cn/kxyj/qwfb/bps/202412/P020241227660032159191.pdf). In the fourth part, "Deepening the Implementation of Collaborative and Agile AI Governance Models," it clearly requires "to coordinate development and security. Agile governance should be used to balance multiple goals. Past growth that sacrificed security should be avoided, and strict regulation that suppresses innovation should also be avoided. Relevant institutions are supported to collaborate in basic innovation, data sharing, application empowerment, and risk prevention."

Although balancing AI innovation and AI governance has received wide attention, how can AI innovation and AI governance be balanced? What are the ways to achieve this balance? What difficulties do existing solutions face? How can these difficulties be resolved? These questions still need further examination.

At present, some solutions have been proposed to balance AI governance and AI innovation. These solutions may achieve multiple forms of collaboration. For example, flexible international organisations can be established to improve adaptability and quick response. Existing ethical laws can also be applied to technology (Ricciardi, 2023). There are also specific ethical governance applications, such as health technology assessment (Carapinha et al., 2024). In addition, the European Union has transformed AI-related ethical principles into a design tool in the form of a checklist. This tool can dynamically evaluate whether AI projects are compatible with ethical principles (Ricciardi, 2023).

In fact, whether based on principle-based solutions or practical solutions, balancing AI governance and AI innovation still faces challenges. For example, there is a lack of clear digital definitions for AI ethical principles. Some scholars have made attempts, such as applying digital principles to different digital technologies (Liu et al., 2024). Based on a review of the current balance solutions, this study proposes a new approach. It uses causal models to enhance decision explainability and to speed up the governance process.

The research reported in this study differs from previous similar work in that it improves the explainability of automated governance and advocates a human-led, machine-assisted governance model. Essentially, it generates risk probabilities and governance recommendations by introducing a risk management framework and existing experience-based solutions.

However, people may be concerned that this automated approach could have negative effects on the labour market, creating new ethical issues. In fact, the approach proposed in this study is designed to support rather than replace human governance staff. It aims to establish an intelligent assessment system that is complementary to humans and does not compete with them.

The paper is organised as follows. The first part introduces the background. The second part reviews existing AI governance and AI innovation approaches and their limitations. The third part proposes an explainable automated review approach. Using a Bayesian network as an example, it discusses a

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

human-assisted AI risk governance framework based on the Bayesian network and explains its rationale and significance. The conclusion summarises the paper and states that the main purpose of the automated assisted governance framework is to improve transparency and accelerate the governance process.

2. Solutions for balancing AI innovation and AI governance and their issues

The relationship between innovation and regulation often shows tension. However, it is crucial to maintain a balance between technological innovation and governance. Too little supervision can lead to the accumulation of risks, while too much supervision may stifle innovation. Therefore, it is necessary to develop a regulatory framework that adapts to technological innovation.

Overall, current solutions on this topic can be mainly divided into two types: principle-based solutions and practical solutions.

2.1. Principle-based solutions: adapting governance to practice

Principle-based solutions for balancing AI governance and AI innovation refer to guidance that includes policies, laws, technical standards (Cantero and Marsden, 2024; Segate, 2024), corporate plans, technical manuals, and academic research programmes. These principle-based solutions generally follow the idea that governance should adapt to practical development. They mainly include two types of adaptation principles: process-based adaptation and top-down adaptation.

2.1.1. Process-based adaptation principle

The process-based adaptation principle refers to consciously adapting AI ethics during the technology design process. This mainly involves the development stage of technical staff. It requires technical staff to balance technology and ethics consciously, and also requires external regulation and incentives for them. External regulation of technical staff in the process mainly includes educational standards, reward and punishment mechanisms, and public accountability.

First is the educational standards. Some scholars suggest embedding ethics into AI teaching to help students reflect on the possible social impacts of AI and to improve moral awareness (Horton et al., 2022; Williams et al., 2023; Zhang et al., 2023). These practices mean that engineers can be taught ethical principles during learning or practical application to prevent them from unknowingly violating basic human rights.

In addition, process-based balance solutions also include rewards and punishments for technical staff. To prevent these regulations from overly restricting technological innovation, the government can introduce bonus incentive policies for companies that prioritise ethics, transparency, and safety. For example, some scholars advocate that government policies for responsible artificial intelligence innovation should include ethical incentives for technical staff, mechanisms to reward whistleblowers, and adaptive incentive programmes (Baldassarre et al., 2025).

Finally, there is public accountability (Loi and Spielkamp, 2021), which reflects democratic governance. Therefore, implementing public accountability can also serve as an external regulation method for technical staff.

It is reasonable to provide internal and external regulation for technical staff because, from their perspective, the importance of balancing innovation and governance is proportional to the level of public reliance on AI. In addition, according to Lawrence Lessig, the Internet creates a form of privilege and makes this privilege possible (Dávila et al., 2023). Some studies also show that algorithmic bias comes from limited data sets and biased algorithm designers (Chen, 2023). Therefore, given that technical staff have the privilege to set network rules and can produce significant impacts, it is necessary to regulate this privilege during the AI design process.

2.1.2. Top-down adaptation principle

The top-down adaptation principle mainly comes from government regulatory proposals. Its balanced approach is that regulatory mechanisms are updated as technology evolves. This adaptability is expressed

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

in two forms: specific regulation and uniform regulation. Specific regulation is based on the analysis of each situation. It is targeted and includes risk-proportionate regulation (Akinsuru and Iseoluwa, 2025), targeted regulation (Ghimire and Edwards, 2024), and interdisciplinary regulation (Bringas et al., 2022; Pfeiffer et al., 2023; Patel et al., 2024).

Uniform regulation has compatibility. First, it involves active participation in global and cross-field dialogue. For example, regulatory agencies continuously reflect, adapt, and improve while cooperating to develop AI guidelines (Akinrinola et al., 2024). This can be considered a complex governance situation or establish a collaborative network that combines government authority with industry expertise, creating highly adaptive and responsive dynamic laws (Walter, 2024). Second, it provides unified or sustainable solutions. For example, it may suggest achieving uniform international AI regulations, emphasising public participation and international cooperation (Kashefi et al., 2024), building sustainable AI regulations (Hacker, 2024), or adopting a joint AI governance approach that places bioethical principles at the core of AI rules (Fessenko and Jasperse, 2025).

Although these principles highlight the key issue that AI governance should be flexible and adaptive, the characteristics of such technology governance and AI supervision solutions are mainly based on traditional human governance. They may still face problems such as long governance cycles, neglect of key issues, delays, and a lack of dynamic tracking. This is determined by the long governance cycles and the uncertainty of risks brought by emerging technologies.

2.2. Practical solutions: Ethically engineered balance solutions

Balance solutions for governance and innovation to achieve a shift from principles to tools. Unlike principle-based solutions for balancing AI innovation and AI governance, ethically engineered solutions use technical means to balance AI innovation and AI governance. They are also called computational ethics (Poszler et al., 2025) and emphasise incorporating ethical principles into the engineering design process.

Ethically engineered solutions have experienced a shift from the "need for ethical engineering" to the "application of ethical engineering." For example, the Belmont Report, proposed in the late 1970s, promoted the development of medical ethics. In recent years, to maintain a balance between innovation and moral responsibility, a hybrid technical solution has been proposed in the medical field. It provides additional protective technologies, including blockchain, federated learning, homomorphic encryption, and differential privacy. Differential privacy was originally introduced by Reza Shokri and Vitaly Shmatikov (Yanamala et al., 2024). Similar research has also been applied in other fields with related ethical requirements.

The role of ethically engineered balance solutions is to address ethical issues from the source. Engineers who lack ethical consideration may face damage to their own interests and social condemnation. Historical cases show that ignoring engineering ethics can lead to scandals. For example, the tragedy of the Boeing 737 MAX crashes revealed the impact of a culture that prioritises profit over safety (Chary, 2024). Since ethical governance issues can be traced back to engineering design, it is crucial for engineers to predict the ethical impacts of emerging technologies because this can serve as an effective preventive measure. Scholars such as Janna van Grunsven have suggested that responsible anticipation should be cultivated in engineering ethics education. Furthermore, some scholars argue that engineering ethics must involve the field of empirical moral psychology (Clancy and Gammon, 2021). In addition, some countries and institutions have gradually attached importance to engineering ethics. For example, the American Society of Civil Engineers, the American Society of Mechanical Engineers, and the National Society of Professional Engineers all place the moral principle of serving the public first. Engineers themselves also called for the protection of public safety, health, and welfare (Smith et al., 2021) after World War II. Engineers Canada has also proposed Professional Engineering Codes of Ethics (Randall and Strong, 2021).

Based on current practical solutions for ethically engineered approaches, they mainly include embedded ethics (McLennan et al., 2022) and responsible design (Lu et al., 2022).

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

Embedded ethics refers to integrating ethics into a specific area of AI development, focusing on ethical technical research. For example, it involves integrating technical strategies in fields such as healthcare (Tahir et al., 2024), finance (Joseph, 2024), and communication, including fairness-aware algorithms (Radanliev, 2025). Currently, there are methods addressing privacy, fairness, and transparency. For example, Bohdan Kovalevskyi proposed a synthetic alignment method to maintain core moral principles, allowing AI to remain flexible across different cultural and application contexts (Kovalevskyi, 2024).

Responsible design refers to implementing ethical safeguards at all stages of a project life cycle. It focuses on ethical planning throughout the entire engineering design process. Its key point is to strengthen systems of auditability, traceability, and reviewability, and it emphasises supervision, accountability, and transparency (Burr and Leslie, 2023). It includes correcting algorithmic bias. For example, the AI design ethics method (EbD-AI) adopted by the European Commission as part of its ethical review procedures specifies recommended processes and actions for different stages of AI development (Brey and Dainow, 2024).

Although ethically engineered approaches have achieved initial results, the practice of engineering ethics is considered inflexible, highly moral-focused, complex, and slow to implement. In addition, this implementation approach currently faces the following challenges.

2.2.1. The conflict between abstract principle requirements and specific practical applications

This mainly appears in two aspects. First, there are difficulties in converting abstract principles into specific applications in terms of completeness, translatability, and explainability. One, embedded ethics is practice-oriented rather than normative and may have some incompleteness (Bleher and Braun, 2023). Two, it is questioned whether morality can be applied in programming. Three, governance principles and technical practices lack a communication bridge, which leads to few ethical explanations of technology or inaccurate translation of abstract principles into specific practice. Therefore, training interdisciplinary communicators as translation bridges for ethical engineering may help resolve this conflict.

Second, there is no unified standard for guiding principles in ethical engineering practice. On the one hand, this is because moral concepts themselves are uncertain. As Irina Shklovski stated, ethical methods cannot be finalised. Efforts in AI ethics do not end with a clear checklist or a set of principles. They must continuously address new concerns as part of the challenging task of cultivating certainty (Shklovski and Némethy, 2023). This means that ethical concepts are still changing. On the other hand, many ethical concepts are vague, contradictory, and yet similar in meaning. For example, respecting "openness" may conflict with respecting "privacy" (Kazim and Koshiyama, 2021). The meaning of "explainability" and how to implement it is also unclear (Prem, 2025). These uncertainties increase the difficulty of converting principles into practice.

Therefore, it is necessary to ease the conflict between principles and practice. This is a prerequisite for guiding practice. In this regard, governments should further develop clear communication channels and decision-making protocols to propose appropriate ways to mitigate this conflict.

2.2.2. Increased technical costs

Even if the conflict between the principles and technical practices is resolved, there may still be an issue of higher technical costs (Akash et al., 2024). This is because additional technical work is needed. This will increase both the time and financial costs of implementing the technology. For example, the costs of updating and maintaining tools, testing for vulnerabilities by external auditors, salaries, benefits, and training for technical staff, fixing vulnerabilities, upgrading and improving systems, and managing risks (Das et al., 2023).

2.2.3. Reduced technical efficiency

For example, consider differential privacy technology. Differential privacy technology also faces some technical problems. This technology usually needs to balance utility and privacy. Generally, the two are inversely related. If accuracy increases, privacy protection decreases, and the opposite is also true.

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

Although it helps prevent linkage attacks and protect privacy, it lacks effective sensitivity values and may affect the reliability of results (Husnoo et al., 2021).

It can be seen that although embedding governance principles in the engineering design process is necessary and helps to address the possible delay of principle-based schemes, the ethical engineering approach still faces a fundamental challenge. This challenge is how to balance improving efficiency with embedding AI governance values and principles. Therefore, a neutral approach between technology developers and government officials is needed. This approach needs to address: (1) the delay of AI governance principle schemes, and (2) the efficiency loss caused by considering AI governance in the engineering field.

3. Explainable automated review scheme

Overall, the limitations of principle-based and practice-based schemes are mainly reflected in three key words: time, speed, and cost. Focusing on time can also help improve governance speed and reduce the cost of innovation and governance. Therefore, this article attempts to propose a governance approach that shortens the time cycle to balance AI innovation and AI governance. Using automated tools is an important way to reduce the time cycle. In August 2024, the State Council of China issued the Opinions on Deepening the Implementation of the "Artificial Intelligence Plus" Action (https://www.gov.cn/zhengce/zhengceku/202508/content_7037862.htm). It clearly proposed the concept of "AI+ governance capability." In Part Two, Article Five, and Part Three, Article Fourteen, it stated: "Create a new vision of human-machine symbiosis in social governance." "Build a new model of government services that can accurately identify needs, plan services proactively, and manage the whole process intelligently." "Create a new pattern of safe governance with multiple stakeholders." "Promote a public safety governance system for natural persons, digital humans, intelligent robots, and other entities." "Enhance AI-driven monitoring, prediction, simulation, and problem-solving capabilities, and promote intelligent collaborative and precise governance." "Strengthen government guidance and industry self-discipline, insist on inclusiveness and caution, classify and grade governance, and accelerate the formation of a dynamic, agile, and multi-stakeholder AI governance system."

The automated governance scheme is called a technical governance system. Automated governance tools are currently in the development stage. They are considered useful political tools because they can improve efficiency, save costs, and identify bias (Ajunwa, 2022).

Currently proposed automated governance schemes include ethics-based auditing (EBA) (Mökander et al., 2021) and the use of reinforcement learning to dynamically learn moral rules (Liu et al., 2024). These automated governance models create conditions for human-machine collaborative governance. However, the black-box problem of artificial neural networks has become a widely recognised technical issue. This raises questions of responsibility in governance. The moral guidelines for artificial intelligence require specificity, clarity, and transparency. This creates the need for an explainable intelligent framework.

This article focuses on using causal models. It explains where they can be applied in governance and what specific processes their implementation will include. It not only describes the connection between principles and automated tools. Here, the emphasis is on using causal models to support regulation, and they can be applied together with other automated models.

First, this is proposed on the premise of acknowledging the limitations of ethical engineering practices. It does not ignore or fully solve these problems. However, shortening the risk review cycle can partly ease the difficulties. Unlike the principle and practice schemes mentioned above, this approach embeds technical models in the governance process. Here, a causal model, the Bayesian network, is used as an example. AI governance laws and regulations are embedded in it to achieve dynamic supervision. Here, the causal model serves as an auxiliary tool for AI governance.

3.1. Considering Bayesian networks as a tool for AI regulation

A Bayesian network is a belief model that can infer the probability of unknown events based on known information. It can also adjust or update its structure according to new data. By building conditional or

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

causal probability relationships between existing variables, it can reason, predict, and help decide the best options. Bayesian networks can model classification, prediction, decision-making, and data completion. In particular, dynamic Bayesian networks can represent time series data, which means they can make timely and dynamic updates.

In recent years, Bayesian networks have been widely used for risk assessment in different industries (Khan et al., 2021; Wang et al., 2025), such as healthcare (Dritsas and Trigka, 2022, climate (Zhang et al., 2025), energy and finance (Wei et al., 2023), and reliability prediction in aircraft operations (Küçüker and Yet, 2023). They generally use mathematical models and expert evaluations to identify the influencing factors. They are considered effective in extracting the relationships between risk factors, identifying different risk levels, and achieving higher accuracy (You et al., 2025).

There are two main reasons why it can be used for AI governance. First, there are existing laws and regulations that classify AI risks. Second, there is some governance experience. These factors create conditions for applying Bayesian networks to AI governance. In addition, Bayesian networks are suitable for AI governance because they are explainable, dynamic, and can be automated. They can trace responsibility and perform causal reasoning according to risk levels in laws and regulations. They can also adjust probability estimates and update decision recommendations based on new factors, showing dynamic adaptability. More importantly, they can speed up the governance cycle. From a time perspective, these three features of Bayesian networks meet three current needs in AI governance: (1) governance needs a solid basis; (2) governance needs to adjust dynamically with technological innovation; and (3) governance needs to speed up the cycle to address potential delays.

In general, applying Bayesian networks to AI governance has four advantages: (1) Bayesian networks have causal traceability, and their predictions are reliable; (2) probability assessments help support decision-making by regulators; (3) Bayesian networks can update dynamically to meet the requirements of adaptability and dynamism in governance; and (4) using existing AI laws and regulations as input makes the approach compliant.

However, it is acknowledged that some parts cannot be explained by machines and that there is regulatory uncertainty. As experts improve AI governance principles and quantify evaluation details, the automated governance model will also improve.

3.2. Where to use: The preceding stage of applying Bayesian networks in AI governance

Given the reasoning and prediction functions of Bayesian networks, they can serve as a buffer in this context.

They are used in the stage before the regulatory sandbox. Their main role is to provide risk probability assessment, real-time screening, and decision recommendations. This helps in handling risks in a targeted way.

Its specific practical scheme is shown in Figure 1.

Figure 1 shows the basic process of automated supervision based on Bayesian networks. It specifically includes the following steps:

The Bayesian network estimates the probability of risks in emerging technologies based on prior knowledge. On the one hand, if the assessment result is positive, it indicates that the model is considered risky according to the Bayesian network. After expressing the specific risk level externally, the tested model is transferred to the regulatory sandbox. In the sandbox, the model undergoes a second risk verification using a series of known risk detection methods configured by the government and relevant regulatory agencies. If the sandbox assessment shows a true positive, it means the technology has a risk, and the sandbox result matches the Bayesian network prediction, so corrective measures are needed. If the sandbox assessment shows a false positive, it means the technology currently has no risk. This also indicates that the Bayesian network has a high false positive rate, and the tested model sample should be used as a special noise example to help optimise the Bayesian network.

On the other hand, if the Bayesian network assesses the risk of an emerging artificial intelligence technology as negative, it means the model passes the buffer evaluation and can temporarily enter social

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

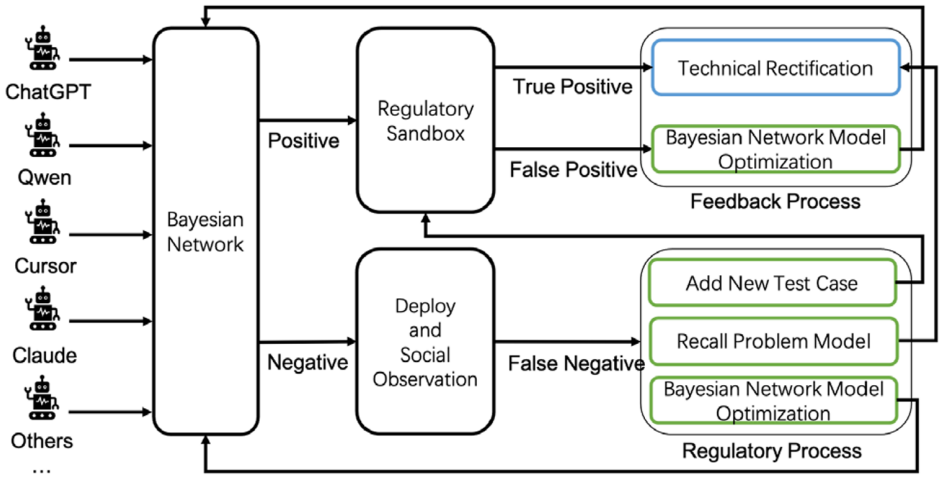


Figure 1. *Automated supervision framework based on Bayesian networks.*

use. This requires the Bayesian network to be highly accurate. However, if the tested model is deployed in society and produces feedback indicating a risk, it means the model is a false negative, and the Bayesian network's previous buffer evaluation was inaccurate. If the model's stability is guaranteed, this may be caused by an unknown new risk type that is not included in the prior knowledge of the Bayesian network. In this case, the new risk definition must be added to the Bayesian network to optimise the model. The problem model should be recalled from the social environment for technical correction, and probability conditions should be updated based on new experience. At the same time, regulatory authorities and other agencies need to coordinate with the model provider to carry out the technical correction and update the Bayesian network's prior knowledge according to the new risk type. The new unknown risk type should also be shared with regulatory authorities and included in sandbox test cases to prepare for other models that may be identified as positive by the Bayesian network in the future. These measures help improve the precision and recall of the entire processing chain.

For the government, using the Bayesian network for risk assessment allows regulators to carry out targeted interventions and improve regulatory efficiency. At the same time, applying the Bayesian network in AI governance together with the regulatory sandbox and full-cycle staff oversight provides a triple guarantee for AI governance. For companies, quickly identifying problems and managing them helps new technologies enter the market faster. This also partially offsets the cost losses caused by ethical engineering and long regulatory cycles.

3.3. How to use: Automated risk assessment and decision recommendation with Bayesian networks

In practical use, the regulatory approach of Bayesian networks is as follows. The "qualitative risk levels" in regulations, such as unacceptable risk, high risk, limited risk, and lowest risk, are converted into nodes. The "experience thresholds" accumulated by regulatory authorities during inspections, such as pass rates, are quantified. These two factors are used as the basis for probability distributions to infer the overall risk level of AI.

The use of Bayesian networks to assist AI governance includes, but is not limited to, the following steps:

(1) Collect legal and regulatory bases for AI governance and extract qualitative criteria for risk assessment.
(2) Convert the regulatory authority's experience, such as pass rates and violation frequencies, into probability tables.

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

(3) Reflect the legal rules and the experience thresholds from step 2 in the causal relationships of the Bayesian network nodes.
(4) Use the Bayesian network to perform probability inference and risk assessment. For example, given a probability value, infer which risk level the AI belongs to and provide corresponding decision suggestions based on the input laws and regulations.

Overall, this model consists of three steps: identification, evaluation, and decision.

First, in the identification stage, in recent years, some scholars have proposed quantitative schemes for ethical governance. For example, Ming Yang combined the EU Artificial Intelligence Act to propose a risk quantification method based on a risk identification matrix. This method can support risk quantification and governance for generative AI (Yang et al., 2025). Xiongliang Zeng and others (https://aiig.tsinghua.edu.cn/__local/4/D1/C8/34D4B0D9D06200E10A0BD14D2BE_F4066969_B4F47.pdf) described a formula for risk level calculation: Risk Level (R) = Risk Probability (P) * Risk Severity (L). They proposed a method to measure risk probability, which involves parameters from historical data, the completeness of safety measures, and the quality of datasets. Risk severity relates to factors such as user scale, amount of property loss, and the scope of social impact. Novelli et al. (2024) proposed a semi-quantitative method under the framework of the Artificial Intelligence Act (AIA) and international security agreements, based on risk scenarios and proportional quantitative assessment. The progress in risk-based quantitative evaluation provides important criteria for AI risk identification and also improves the transparency of the assessment.

Second, in the assessment stage, reference can be made to the EU Artificial Intelligence Act. This Act divides AI systems into four risk levels: (1) minimal risk; (2) limited risk; (3) high risk; and (4) unacceptable risk. This classification can avoid a one-size-fits-all approach to management.

The minimum risk has no obligation restrictions, such as spam and filtering. Limited risk is related to manipulation and deception. High risk is related to the health, safety, and fundamental rights of natural persons and to environmental impacts. Unacceptable risk is related to human fundamental rights, and its use must be prohibited.

Finally, in the decision stage, some scholars have given decision recommendations for different risks. Szadeczky and Bederna (2025) determined different risk management strategies for AI systems, ranging from minimal to high risk based on legal classifications. For minimal risk, general principles, such as human supervision, non-discrimination, and fairness, should be followed. For limited risk, transparency requirements and the provision of necessary labels should be applied. Unacceptable risks should be prohibited. High risks should follow the EU Artificial Intelligence Act, which sets obligations for high-risk systems, including risk identification, testing, and risk mitigation measures. During model training, data governance and management practices should be applied, technical documents drafted and updated, and automated logs maintained (Act EAI, 2024). These measures provide a basis or reference for machine decisions.

For risks that cannot be measured and rely heavily on subjective judgement, the standards for risk assessment can be jointly developed by legal, ethical, and technical experts.

For example, AI that manipulates people using age or the vulnerability of disabled persons is considered an unacceptable risk and must be banned from the market and fined. For high-risk medical diagnostic systems, recommendations include system verification, testing, human review, professional certification, and record keeping. AI, when judged as limited risk, such as AI-generated content, must provide the necessary source information. For example, a game recommendation system can be considered low risk and may be freely used under the Data Protection Act.

For example, if a causal model identifies age discrimination in a company's use of AI for recruitment, the system will assign a high-risk level and provide corresponding governance advice. This advice may include recommendations based on laws, such as the Employment Age Discrimination Act, or on existing experience, including fines.

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

3.4. Framework construction and validation

3.4.1. Defining risk variables based on the EU AI Act

The table is shown in Table 1.

3.4.2. Taking an intelligent recruitment screening system as an example

This example strictly follows the compliance requirements of the EU AI Act. It presents the full process from legal rule analysis and numerical risk reasoning to the automatic generation of governance

Table 1. *Computable variables and regulatory measures mapping table*

Computable variable	Indicator (example)	Related rule	Regulatory measure
Manipulation	Whether the system influences behaviour by using cognitive bias	Art. 5	Ban on release or removal from the market, fines, mandatory correction
Vulnerability exploitation	Whether the system targets children or disabled groups	Art. 5	Prohibition, compliance review
Severe rights impact	Whether the system directly harms equality rights or personal rights	Art. 5	Prohibition, legal responsibility
Health and safety impact	Accident rate, misdiagnosis rate, failure rate	Art. 15	Safety testing, robustness assessment, fail safe protection
Rights impact	Employment or credit discrimination impact, complaint rate	Art. 9	Risk assessment, rights impact assessment
System risk	Failure rate, success rate of adversarial attacks	Art. 15	Security strengthening, red team testing
Scale	Number of affected people normalised	Annex III (indirect)	Graded deployment, staged release
Severity	Possible property or rights loss	Art. 9	Preventive control, emergency plan
Data bias	Fairness difference, suitability, completeness, equality	Art. 10	Bias reduction, reweighting, fairness constraints
Data quality	Relevance, originality, availability	Art. 10	Data cleaning, label review
Representativeness	Insufficient coverage index	Art. 10	Expanded sampling, stratified sampling
Transparency	Disclosure coverage, including whether AI involvement and use limits are explained	Art. 13	Additional instructions for use, mechanism description
Explainability	Ratio of explainable outputs or explanation success rate	Art. 13	Explanation interface, reports
Human oversight	Ratio of human review or ability to stop the system	Art. 14	Threshold-based review, one-click stop
Logging	Coverage rate of key event records	Art. 12	Full logging, audit interface
Documentation	Completeness rate of documents	Art.11/ Annex IV	Technical documents, update records
Risk management	Risk assessment frequency or completion rate	Art. 9	Establish a continuous risk management system

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

suggestions. Relevant calculation steps are also provided to ensure that the results can be reproduced and verified.

First, it is necessary to determine the basic condition for starting the whole risk assessment process. According to Annex III of the EU AI Act, AI systems used for recruitment, staff screening, and promotion are clearly defined as high-risk systems. Therefore, the risk assessment needs to consider risks related to discrimination, non-discrimination, and fairness. More specifically, requirements need to be set for the quality of training, validation, and test datasets. This includes checking historical bias, ensuring data representativeness, adding suitable human review procedures, and taking necessary corrective measures.

Second, the Bayesian network nodes and their parameters used in this example are given in Table 2.

Among these nodes, *HB*, *DR*, and *BM* are root nodes. Their prior probabilities are set based on common bias problems in the industry. *DO* and *HR* are non-root nodes. Their probabilities are determined by their parent nodes and are quantified through conditional probability tables (CPTs). For example, the CPT of the intermediate node "Discriminatory Outcome" is set as follows. When $HB = High$, $DR = Poor$, and $BM = Ineffective$, the probability of a discriminatory outcome is the highest and is set at 0.95. When all three conditions are in ideal states, the probability is the lowest and is set at 0.10. This setting reflects the real situation in which multiple factors together can sharply increase risk.

Finally, conduct a prior risk assessment. Under the default condition, it is necessary to assess the possibility that this intelligent recruitment system may contain potential risks. The state of *HR* is determined only by *DO*. Therefore, the prior probability of the intermediate node "Discriminatory Outcome," namely $P(DO = Likely)$, is first calculated. Then, $P(HR = Yes)$ is calculated based on this result.

Since *DO* has three parent nodes, namely *HB*, *DR*, and *BM*, and each node has two states, there are $2^3 = 8$ possible combinations. It is necessary to examine all combinations, calculate the probability of each combination, and calculate the conditional probability that *DO* is Likely under each combination. After that, all results are added together. The calculation process is as Table 3.

Thus, we obtain the prior probability that the system produces a discriminatory outcome without additional information as:

$$P(DO = Likely) \times \Sigma P(DO = Likely | HB, DR, BM) \times P(HB) \times P(DR) \times P(BM) = 0.5725$$

Therefore, in the initial state, the recruitment system has a 57.25% probability of producing a discriminatory outcome.

Table 2. *Bayesian network nodes and their parameters*

Risk node	State definition	AIA legal rule	Prior probability
Historical data bias (HB)	High/Low	Art. 10: Check historical data bias	P(HB=High) = 0.60
Data representativeness (DR)	Poor/Good	Art.10:Data should represent the target group	P(DR = Poor) = 0.70
Bias mitigation effectiveness (BM)	Ineffective/Effective	Art. 10: Appropriate bias mitigation measures should be taken	P(BM = Ineffective) = 0.50
Discriminatory outcome (DO)	Likely/ Unlikely	Art. 9: Risk assessment should consider discrimination risk	Defined by CPT
High-risk determination (HR)	Yes/No	Art. 6 and Annex III: High-risk system determination standard	Defined by CPT

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

Table 3. *The calculation process of the prior probability that the intelligent recruitment screening system produces discriminatory results*

Combo ID	HB	DR	BM	$P(Combo)$	$P(DO = Likely \mid Combo)$	Contribution
1	High	Poor	Ineffective	$0.6 \times 0.7 \times 0.5 = 0.21$	0.95	$0.21 \times 0.95 = 0.1995$
2	High	Poor	Effective	$0.6 \times 0.7 \times 0.5 = 0.21$	0.70	$0.21 \times 0.70 = 0.1470$
3	High	Good	Ineffective	$0.6 \times 0.3 \times 0.5 = 0.09$	0.60	$0.09 \times 0.60 = 0.0540$
4	High	Good	Effective	$0.6 \times 0.3 \times 0.5 = 0.09$	0.40	$0.09 \times 0.40 = 0.0360$
5	Low	Poor	Ineffective	$0.4 \times 0.7 \times 0.5 = 0.14$	0.50	$0.14 \times 0.50 = 0.0700$
6	Low	Poor	Effective	$0.4 \times 0.7 \times 0.5 = 0.14$	0.30	$0.14 \times 0.30 = 0.0420$
7	Low	Good	Ineffective	$0.4 \times 0.3 \times 0.5 = 0.06$	0.30	$0.06 \times 0.30 = 0.0180$
8	Low	Good	Effective	$0.4 \times 0.3 \times 0.5 = 0.06$	0.10	$0.06 \times 0.10 = 0.0060$
Total	–	–	–	1.00	–	$\Sigma = 0.5725$

Next, we calculate the prior probability of the top level node, which is the high-risk decision P (HR = Yes). The same law of total probability is used:

$$P(HR = Yes) = P(HR = Yes|DO = Likely) \times P(DO = Likely) + P(HR = Yes|DO = Unlikely) \times P(DO = Unlikely)$$

Substituting the known values, where $P(HR = Yes|DO = Likely) = 0.90$, $P(HR = Yes|DO = Unlikely) = 0.10$, and $P(DO = Unlikely) = 1 - P(DO = Likely) = 1 - 0.5725 = 0.4275$, we have:

$$P(HR = Yes) = 0.90 \times 0.5725 + 0.10 \times 0.4275 = 0.51525 + 0.04275 = 0.5580$$

From this result, we obtain that without any additional evidence, the prior probability that the system is classified as "high risk" is 55.8%. This value is already higher than most internal control thresholds in companies, for example, 50%. It shows that the system has a clear compliance risk even before deployment.

3.4.3. Dynamic update of risk probability

The following section demonstrates how to use dynamic evidence to drive Bayesian updating in the model and then carry out risk attribution analysis on this basis.

(1) Assume that a third-party organisation conducts a bias audit on the recruitment system and finds serious issues. The specific evidence E is that the four-fifths rule is not satisfied. The hiring rate for female candidates compared with male candidates is 61.5%. This is below the legal threshold of 80%. This value is the key input that drives the model update.

(2) The likelihood function is set. It describes the probability of observing the evidence under different states of "Discriminatory Outcome" as $P(E|DO = Likely) = 0.92$. If the system does have a discriminatory outcome, then the audit has a 92% chance to detect this bias: $P(E|DO = Unlikely) = 0.15$. If the system does not have a discriminatory outcome, the audit still has a 15% chance of giving a false positive or reporting bias. This non-zero probability takes into account possible sampling errors or other limits in the audit process.

(3) Bayesian update. We calculate the posterior probability $P(DO = Likely|E)$. We use Bayes' theorem to compute the posterior probability of "Discriminatory Outcome" given the observed audit failure evidence E. The formula of Bayes' theorem is as follows:

$$P(DO = Likely|E) = \frac{P(E|DO = Likely) \times P(DO = Likely)}{P(E)}$$

In this equation, the denominator $P(E)$ is the total probability of observing the evidence E. It can be calculated using the law of total probability:

$$P(E) = P(E|DO = Likely) \times P(DO = Likely) + P(E|DO = Unlikely) \times P(DO = Unlikely)$$

Substituting the values:

$$P(E) = 0.92 \times 0.5725 + 0.15 \times (1 - 0.5725) = 0.5267 + 0.064125 = 0.590825$$

The posterior probability is:

$$P(DO = Likely|E) = \frac{0.92 \times 0.5725}{0.590825} \approx 0.8915$$

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

After the impact of this single piece of evidence from the audit failure, the prior probability of 57.25% for a discriminatory outcome is significantly increased to a posterior probability of 89.15%.

3.4.4. Reverse inference: Identifying key risk sources $P(HB = High|E)$

We conduct a reverse inference to calculate the posterior probability that the root node "Historical Data Bias" is in the state High under the evidence of audit failure. This is written as $P(HB = High|E)$. This step is used for risk attribution.

Here, we calculate the contribution weight of key combinations. For example, $Combo_1(HB = High, DR = Poor, BM = Ineffective)$ contributes to the explanation of evidence E as follows:

$$\begin{aligned} Contrib_1 &= P(E|DO = Likely) \times P(DO = Likely|Combo_1) \times P(Combo_1) \\ &= 0.92 \times 0.95 \times 0.21 = 0.18354 \end{aligned}$$

The weight of this combination in the total probability of evidence $P(E)$ is:

$$Weight_1 = \frac{Contrib_1}{P(E)} = \frac{0.18354}{0.590825} \approx 0.3107$$

Using the same method, we calculate the weights of all combinations where $HB = High$, which include $Combo_1$, $Combo_2$, $Combo_3$, and $Combo_4$. We then sum these values. The total contribution of $HB = High$ is about 69.5%. This value is much higher than its prior probability of 60%.

Attribution conclusion: under the audit failure evidence, historical data bias is the main cause of the current high risk state.

$$\begin{aligned} P(HR = Yes|E) &= P(HR = Yes|DO = Likely) \times P(DO = Likely|E) \\ &+ P(HR = Yes|DO = Unlikely) \times P(DO = Unlikely|E) \\ &= 0.90 \times 0.8915 + (1 - 0.9) \times (1 - 0.8915) = 0.8132 \end{aligned}$$

The probability of high-risk classification increases from 55.8% to 81.32%. This clearly triggers the mandatory requirement in Article 16 of the EU AI Act that corrective measures must be taken for high-risk systems.

3.4.5. Generation of automated governance recommendations

We can build a simple rule engine. This engine maps the numerical outputs of the model directly to specific governance actions and related legal provisions. The process is shown in Table 4.

3.4.6. Explanation output

Table 5 is an example of a compliance risk summary designed for regulators.

Recommended actions: (1) Immediately suspend the automatic screening function (based on Art. 16). (2) Submit a bias root cause analysis report within 72 hours (based on Art. 17). (3) Retraining must pass review by an independent ethics committee before it starts (based on Art. 9).

3.4.7. Continuous iteration and governance

The results of governance actions can be used as new evidence. This feedback is sent back into the network and drives model iteration and development. For example, if the company adopts the recommendation and takes the following actions:

(1) Data resampling is completed. This means that the situation of historical data bias and data representativeness is improved. We update $P(DR = Good)$ from 0.3 to 0.7.

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

Table 4. *Automated governance recommendations*

Trigger condition	Corresponding AIA article	Governance recommendation
$P(HR = Yes \mid E) > 80\%$	Art. 16	Immediately stop the automatic decision function of the system in the production environment until the correction work is completed.
$P(HB = High \mid E) > 65\%$	Art. 10	Launch a special audit on bias in historical data. The focus is to analyse differences in hiring rates across dimensions such as gender and age.
$P(BM = Ineffective \mid E) > 50\%$	Art. 10	Redesign the bias mitigation plan. Introduce adversarial debiasing techniques to improve the effectiveness of the mitigation measures.
$P(DO = Likely \mid E) > 85\%$	Art. 14	Enable a mandatory human review routing system. All low confidence predictions must be manually checked and confirmed by the human resources department.

Table 5. *Example of a compliance risk summary*

Item	Content
Risk level	High risk (confidence: 81.3%)
Main cause	Historical data bias (contribution: 69.5%)
Evidence support	Third-party audit found that the four-fifths rule is not satisfied. The female hiring rate is 61.5%, which is below the 80% threshold.
Legal link	Violation of data governance requirements under Art. 10(2)(g) of the EU AI Act

(2) An adversarial debiasing algorithm is deployed. This improves the effectiveness of bias mitigation measures. We update $P(BM = Effective)$ from 0.5 to 0.8.

Next, we input the updated prior probabilities into the network and run the inference process again. The new calculation will show whether $P(HR = Yes)$ has fallen to an acceptable level, for example below 30%. If the risk is reduced, it shows that the governance measures are effective. If the risk is still high, it is necessary to find new evidence, such as new monitoring logs, and apply deeper governance actions. This forms a spiral and continuous risk management loop. This mechanism ensures that compliance of the AI system is not static. It is a dynamic ability that can adapt to changes in internal and external environments.

Therefore, the Bayesian network, together with dynamic evidence-based probability updates and an automatic rule engine, can convert the abstract legal requirements of the EU AI Act into a risk governance framework for an intelligent recruitment system. This framework is quantifiable, explainable, and executable.

3.5. The significance of an interpretable and adaptive AI governance framework

The interpretable and adaptive AI governance framework based on Bayesian network technology has more advantages than principle-based and practical schemes in shortening the cycle of AI innovation and AI governance. It also performs better in interpretability than other automated governance approaches. In addition, the balanced significance of the interpretable and adaptive AI governance framework is reflected

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

in several aspects. First, the Bayesian network model has general applicability, which helps provide an inclusive governance path. Second, the automatic updating ability of the Bayesian network helps to compensate for current static governance strategies and supports a combination of dynamic and static governance. Third, the Bayesian network can perform reverse reasoning, which helps to automatically generate value alignment schemes.

3.5.1. The interpretable and adaptive AI governance framework helps to provide an inclusive governance path

At present, there is a global competition in artificial intelligence regulation (Neuwirth, 2024). However, balancing AI governance and AI innovation is a common challenge faced worldwide. The interpretable and adaptive AI governance framework has general applicability.

First, the interpretable and adaptive AI governance framework is scalable and can be applied in different countries. Different national risk quantification schemes can be entered into causal models, such as Bayesian models, making it possible to promote their use across countries. Second, the interpretable and adaptive AI governance framework can balance AI innovation and AI governance. On one hand, companies that follow this framework can shorten the time for technology release and application, which helps to avoid profit loss caused by long review periods. On the other hand, governments can use this framework to address the potential delay and static nature of traditional governance schemes.

This requires:

(1) active dialogue between countries to discuss and establish reasonable quantitative standards for artificial intelligence risk assessment; (2) regulators and experts providing quantitative standards for ethical and compliance judgements to technical personnel; and (3) technical personnel actively understanding the relevant quantitative standards and embedding them into causal models.

In conclusion, this depends on cross-disciplinary cooperation, mainly involving technical staff, evaluators (experts), and other stakeholders. The purpose is to promote the optimisation of relevant causal models, especially in dealing with new ethical situations. A well-trained causal model can infer potential or similar ethical causes based on risk assessment results. This can reduce potential ethical problems that may arise from AI innovation.

Therefore, the explainable adaptive AI governance framework has the advantage of universality and can serve as an inclusive governance approach.

3.5.2. The explainable adaptive AI governance framework can serve as a supplement to static governance strategies

AI dynamic governance and static governance can complement each other. Because technology changes quickly, dynamic governance is necessary. Since regulatory uncertainty and ethical rules are hard to formalise, governance should combine static and dynamic strategies for the long term. Expert assessment should be the main approach, supported by machines.

This requires: (1) regulators to develop multiple dynamic governance plans, for example, combining stage-based supervision with full-cycle supervision; (2) technical staff to develop reliable assessment and decision tools for different purposes; and (3) companies to conduct self-inspections according to laws and regulations.

3.5.3. The explainable adaptive AI governance scheme helps to automatically generate value alignment strategies

The explainable adaptive AI governance scheme produces risk probability values and risk types that can generate appropriate value-based decision suggestions. For example, when the causal model shows that a technology has a high risk of opacity, it will suggest increasing transparency and seeking decision suggestions from the decision database that align with values. For this, it is essential to input complete value strategies into the causal model to provide correct suggestions. In addition, when new technologies may cause ethical issues that go beyond the existing decision database, experts are needed to assess and

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

analyse the new ethical problems, provide new decision suggestions, and input them into the causal model.

This requires: (1) experts to continuously predict and publish potential ethical issues in AI development; (2) technical staff to update the causal model structure in time according to expert suggestions; and (3) regulators to act as a bridge by classifying existing ethical decisions and promptly communicating new ethical decisions to companies.

4. Conclusion

Through the work in this article, we provide an overall view of current approaches to balancing AI governance and AI innovation.

First, the principle-based approach to balancing AI governance and AI innovation proposes flexibility and adaptability in the regulatory framework, but it may have potential delays. Second, the practice-based approach can embed ethical requirements into the technology development process, but it may come at the cost of harming the interests of enterprises. This article presents an explainable and adaptive AI governance approach, which aims to improve governance transparency and shorten the governance cycle. It reflects both technical rationality and value rationality. Its main features are explainability and adaptability. The significance of the explainable and adaptive AI governance approach is as follows. First, it helps provide an inclusive governance path. Second, it helps compensate for static governance strategies. Third, it helps automatically generate value-aligned plans. It can be seen that this approach is an automated prediction and reasoning framework that supports risk pre-assessment in governance. However, this approach still has some limitations. First, some ethical judgements are hard to quantify or have regulatory uncertainty. Second, optimising the causal model effectively requires time and cost. Given these limitations, the approach should be seen as a tool for AI governance rather than a replacement. It should be noted that this article only takes a step towards a human-machine collaborative governance method. The practical framework of the explainable and adaptive AI governance approach is gradually improving. The causal model is not the only solution, but serves as a supplement to current approaches. Coordinated governance using multiple approaches can better ensure a balance between innovation and governance. The issue of balancing AI innovation and AI governance is a topic that deserves long-term attention.

Data availability statement. This article does not cover data and code.

Author contributions. Y.W.: conceptualisation, data curation, formal analysis, methodology, investigation, validation, visualisation, writing – original draft, writing – review and editing, project administration, resources, supervision.

Funding statement. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. During the preparation of this work, the author used ChatGPT for translation and Grammarly for grammar checking. After using these tools, the author carefully reviewed and edited the content as needed to ensure accuracy and clarity.

Provenance. This article was submitted for consideration for the 2025 Data for Policy Conference to be published in Data and Policy on the strength of the Conference review process.

References

Act EAI (2024) *The Eu Artificial Intelligence Act*. Luxembourg: Publications Office of the European Union.

Ajunwa I (2022) Automated governance. *NCL Review 101*, 355.

Akash TR, **Lessard NDJ**, **Reza NR and Islam MS** (2024) Investigating methods to enhance data privacy in business, especially in sectors like analytics and finance. *Journal of Computer Science and Technology Studies 6*(5), 143–151. https://doi.org/10.32996/jcsts.2024.6.5.12.

Akinrinola O, **Okoye CC**, **Ofodile OC and Ugochukwu CE** (2024) Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews 18*(3), 050–058. https://doi.org/10.30574/gscarr.2024.18.3.0088.

Akinsuru A and Iseoluwa J (2025) *Balancing Innovation and Regulation: A Framework for Technology Governance*. ResearchGate.

Baldassarre MT, **Caivano D**, **Nieto BF**, **Ragone A** (2025) *Ethics-Driven Incentives: Supporting Government Policies for Responsible AI Innovation*. IEEE Intelligent Systems, *40*(4), 55–63. Los Alamitos, CA: IEEE Computer Society. https://doi.org/10.1109/MIS.2025.3583222.

Bleher H and Braun M (2023) Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics 29*(3), 21. https://doi.org/10.1007/s11948-023-00443-3.

Brey P and Dainow B (2024) Ethics by design for artificial intelligence. *AI and Ethics 4*(4), 1265–1277. https://doi.org/10.1007/s43681-023-00330-4.

Bringas CA, **Nannini L**, **Rieger A**, **Scott KM**, **Zhao X**, **Patro GK and Kinder KK** (2022) Fairness in agreement with European values: An interdisciplinary perspective on AI regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 107–118.

Burr C and Leslie D (2023) Ethical assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics 3*(1), 73–98. https://doi.org/10.1007/s43681-022-00178-0.

Cantero GM and Marsden CT (2024) Artificial intelligence co-regulation? The role of standards in the EU AI Act. *International Journal of Law and Information Technology 32*, eaae011.

Carapinha JL, **Botes D and Carapinha R** (2024) Balancing innovation and ethics in AI governance for health technology assessment. *Journal of Medical Economics 27*(1), 754–757.

Chary S (2024) Employee grievance redressal and corporate ethics: Lessons from the Boeing 737-MAX crashes. *Science and Engineering Ethics 30*(2), 14. https://doi.org/10.1007/s11948-024-00475-3.

Chen Z (2023) Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications 10*(1), 1–12. https://doi.org/10.1057/s41599-023-02079-x.

Clancy RF and Gammon A (2021) The ultimate goal of ethics education should be more ethical behaviors. In *2021 ASEE Virtual Annual Conference Content Access*.

Das D, **Banerjee S**, **Chatterjee P and Ghosh U** (2023) A comprehensive analysis of trust, privacy, and security measures in the digital age. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications*, 360–369.

Dávila J, **De FP**, **Lessig L and Perrone I** (2023) ¿ El código sigue siendo ley?: Entrevista con Lawrence Lessig. *Hipertextos 11*(20), 077. https://doi.org/10.24215/23143924e077.

Dritsas E and Trigka M (2022) Data-driven machine-learning methods for diabetes risk prediction. *Sensors 22*(14), 5304. https://doi.org/10.3390/s22145304.

Fessenko DS and Jasperse A (2025) Ethics at the heart of AI regulation. *AI and Ethics 5*(3), 3387–3398. https://doi.org/10.1007/s43681-024-00562-y.

Ghimire A and Edwards J (2024) From guidelines to governance: A study of ai policies in education. In *International Conference on Artificial Intelligence in Education*, 299–307.

Hacker P (2024) Sustainable AI Regulation. *Common Market Law Review 61*(2), 345–386. https://doi.org/10.54648/COLA2024025.

Horton D, **McIlraith SA**, **Wang N**, **Majedi M**, **McClure E and Wald B** (2022) Embedding ethics in computer science courses: Does it work? In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*. *1*, 481–487.

Husnoo MA, **Anwar A**, **Chakrabortty RK**, **Doss R and Ryan MJ** (2021) Differential privacy for IoT-enabled critical infrastructure: A comprehensive survey. *IEEE Access 9*, 153276–153304. https://doi.org/10.1109/ACCESS.2021.3124309.

Joseph S (2024) Balancing data privacy and compliance in blockchain-based financial systems. *Journal of Engineering Research and Reports 26*(9), 169–189. https://doi.org/10.9734/jerr/2024/v26i91271.

Kashefi P, **Kashefi Y and Ghafouri MA** (2024) Shaping the future of AI: Balancing innovation and ethics in global regulation. *Uniform Law Review 29*(3), 524–548. https://doi.org/10.1093/ulr/unae040.

Kazim E and Koshiyama AS (2021) A High-Level Overview of AI Ethics. *Patterns (N Y) 2*(9), 100314. https://doi.org/10.1016/j.patter.2021.100314.

Khan RU, **Yin J**, **Mustafa FS and Anning N** (2021) Risk assessment for berthing of hazardous cargo vessels using Bayesian networks. *Ocean & Coastal Management 210*, 105673. https://doi.org/10.1016/j.ocecoaman.2021.105673.

Kovalevskyi B (2024) Ethics and safety in ai fine-tuning. *Journal of Artificial Intelligence General Science 1*(1), 259–267. https://doi.org/10.60087/jaigs.v1i1.275.

Küçüker FU and Yet B (2023) Reliability prediction for aircraft fleet operators: A Bayesian network model that combines supplier estimates, maintenance data and expert judgement. *Journal of the Operational Research Society 74*(10), 2187–2198. https://doi.org/10.1080/01605682.2022.2129486.

Liu Y, **Zheng W and Su Y** (2024) Enhancing ethical governance of artificial intelligence through dynamic feedback mechanism. In *International Conference on Information*, 105–121.

Loi M and Spielkamp M (2021) Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 757–766

Lu Q, **Zhu L**, **Xu X**, **Whittle J and Xing Z** (2022) Towards a roadmap on software engineering for responsible AI. In *Proceedings of the 1st International Conference on AI Engineering: software Engineering for AI*.

McLennan S, **Fiske A**, **Tigard D**, **Müller R**, **Haddadin S and Buyx A** (2022) Embedded ethics: A proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics 23*(1), 6. https://doi.org/10.1186/s12910-022-00746-3.

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press

Mökander J, **Morley J**, **Taddeo M and Floridi L** (2021) Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics 27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4.

Neuwirth RJ (2024) The global institutional governance of AI: A four-dimensional perspective. *International Journal of Digital Law and Governance 1*(1), 113–153. https://doi.org/10.1515/ijdlg-2024-0004.

Novelli C, **Casolari F**, **Rotolo A**, **Taddeo M and Floridi L** (2024) AI risk assessment: A scenario-based, proportional methodology for the AI act. *Digital Society 3*(1), 13. https://doi.org/10.1007/s44206-024-00095-1.

Patel AU, **Gu Q**, **Esper R**, **Maeser D and Maeser N** (2024) The crucial role of interdisciplinary conferences in advancing explainable AI in healthcare. *BioMedInformatics 4*(2), 1363–1383. https://doi.org/10.3390/biomedinformatics4020075.

Pfeiffer J, **Gutschow J**, **Haas C**, **Möslein F**, **Maspfuhl O**, **Borgers F and Alpsancar S** (2023) Algorithmic fairness in AI. *Business & Information Systems Engineering 65*(2), 209–222. https://doi.org/10.1007/s12599-023-00787-x.

Poszler F, **Portmann E and Lütge C** (2025) Formalizing ethical principles within AI systems: Experts' opinions on why (not) and how to do it. *AI and Ethics 5*(2), 937–965. https://doi.org/10.1007/s43681-024-00425-6.

Prem E (2025) From ethical AI frameworks to tools: A review of approaches. *AI and Ethics 3*(3), 699–716.

Radanliev P (2025) AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence 39*(1), 2463722. https://doi.org/10.1080/08839514.2025.2463722.

Randall EJ, **Strong D** (2021) Understanding engineers' ethical environmental responsibility. In *Proceedings of the Canadian Engineering Education Association*.

Ricciardi CL (2023) The dilemma of rapid AI advancements: Striking a balance between innovation and regulation by pursuing risk-aware value creation. *Information 14*(12), 645. https://doi.org/10.3390/info14120645.

Segate RV (2024) Drafting a cybersecurity standard for outer space missions: On critical infrastructure, China, and the indispensability of a global inclusive approach. *Journal of Asian Security and International Affairs 11*(3), 345–375. https://doi.org/10.1177/23477970241261432.

Shklovski I and Némethy C (2023) Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society 26*(1), 37–53. https://doi.org/10.1080/1369118X.2021.2014547.

Smith NM, **Zhu Q**, **Smith JM and Mitcham C** (2021) Enhancing engineering ethics: Role ethics and corporate social responsibility. *Science and Engineering Ethics 27*(3), 28. https://doi.org/10.1007/s11948-021-00289-7.

Szadeczky T and Bederna Z (2025) Risk, regulation, and governance: Evaluating artificial intelligence across diverse application scenarios: Risk, regulation, and governance. *Security Journal 38*(1), 35. https://doi.org/10.1057/s41284-025-00495-z.

Tahir NUA, **Rashid U**, **Hadi HJ**, **Ahmad N**, **Cao Y**, **Alshara MA and Javed Y** (2024) Blockchain-based healthcare records management framework: Enhancing security, privacy, and interoperability. *Technologies 12*(9), 168. https://doi.org/10.3390/technologies12090168.

Walter Y (2024) Managing the race to the moon: Global policy and governance in artificial intelligence regulation—A contemporary overview and an analysis of socioeconomic consequences. *Discover Artificial Intelligence 4*, 14.

Wang J, **Hu J and Pang L** (2025) An interpretable dynamic Bayesian network method for time-varying seismic liquefaction uplift risk assessment of underground rectangular structures. *Engineering Applications of Artificial Intelligence 159*, 111746. https://doi.org/10.1016/j.engappai.2025.111746.

Wei L, **Yu H and Li B** (2023) Energy financial risk early warning model based on Bayesian network. *Energy Reports 9*, 2300–2309. https://doi.org/10.1016/j.egyr.2022.12.151.

Williams R, **Ali S**, **Devasia N**, **DiPaola D**, **Hong J**, **Kaputsos SP and Breazeal C** (2023) AI+ ethics curricula for middle school youth: Lessons learned from three project-based curricula. *International Journal of Artificial Intelligence in Education 33*(2), 325–383. https://doi.org/10.1007/s40593-022-00298-y.

Yanamala AKY, **Suryadevara S and Kalli VDR** (2024) Balancing innovation and privacy: The intersection of data protection and artificial intelligence. *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence 15*(1), 1–43.

Yang M, **Wang J**, **Zhang T**, **Jiang R and Gao T** (2025) Risk identification, quantification and governance strategies for generative artificial intelligence service. *Chinese Management Studies 20*(5), 1574–1594.

You M, **Pan X and Zhu C** (2025) Risk identification and assessment of internet public opinion on public emergencies based on Bayesian network and association rule mining. *Frontiers in Public Health 13*, 1642960. https://doi.org/10.3389/fpubh.2025.1642960.

Zhang H, **Lee I**, **Ali S**, **DiPaola D**, **Cheng Y and Breazeal C** (2023) Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of Artificial Intelligence in Education 33*(2), 290–324. https://doi.org/10.1007/s40593-022-00293-3.

Zhang W, **Cui J**, **Yao W**, **Guida M**, **Yeboah FK**, **Zhou X and Liu G** (2025) Bayesian network modeling of flood cascade and climate risks in the Pearl River Delta. *NPJ Natural Hazards 2*(1), 68. https://doi.org/10.1038/s44304-025-00115-1.

Cite this article: Wang Y (2026). An interpretable and adaptive AI governance framework. *Data & Policy*, 8: e48. doi:10.1017/dap.2026.10099

https://doi.org/10.1017/dap.2026.10099 Published online by Cambridge University Press