https://doi.org/10.58245/ipsi.tir.2603.03

Validation-Centric AI-Native Decision Systems: Integrating the GAVA Loop with PRIME–INSPECT Governance

Avramović, Nebojša; Čomić, Tijana; Marković, Aleksandar; Čavoški, Sava; Zornić, Nikola; and Vujović, Vladimir

Abstract: AI-native decision systems — in which generative artificial intelligence generates, executes, and adapts decisions in real time — are shifting the centre of gravity of analytical effort. Whereas the principal challenge used to be the extraction of insight, it is now becoming the validation of AI outputs while action is still being taken upon them. Existing approaches address this problem only partially: explainable artificial intelligence and governance frameworks treat validation as an ex post layer; reinforcement learning reduces it to a single scalar reward; control theory and the broader cybernetic tradition formalise the closed loop but neglect its socio-technical constraints; while the PRIME–INSPECT framework establishes a governance foundation, yet leaves real-time validation implicit. This paper proposes a validation-centric architecture that links the GAVA decision loop (Generate–Act–Validate–Adapt), introduced here, with the PRIME–INSPECT framework and elevates validation to a central analytical function comprising four subdimensions: statistical, operational, cognitive, and governance-related. The proposed architecture is empirically examined on a dual sample of IT professionals and top-management representatives, using descriptive statistics, multiple regression, mediation, moderation, and structural equation modelling. The results confirm the central role of trust, the negative effect of perceived risk, and the importance of top-management support, while robustness checks separated by the IT and TMT subsamples further strengthen the structural model. Taken as a whole, the findings support the view that validation should be treated as a first-order organisational stage in AI-native decision systems.

Index Terms: artificial intelligence governance, AI-native systems, behavioural intention, explainable artificial intelligence (XAI), GAVA loop, generative artificial intelligence, real-time decision-making, structural equation modelling, trust calibration, validation-centric architecture

1. Introduction

Artificial intelligence is transforming the way organisations make decisions, with generative artificial intelligence and foundation models [1] at the centre of this transformation. For most of their development, artificial intelligence systems played an auxiliary role: they analysed historical data and provided users with insights on the basis of which those users acted. That relationship is now changing. Large language models, multimodal systems, and autonomous agents [2] have enabled the emergence of a different kind of system — AI-native systems that independently generate, evaluate, and execute decisions in real time [3], [4], [5].

This is a change that goes beyond mere improvement. In data-driven environments, the main constraints were data availability and the ability to extract meaningful insights. In AI-native systems, the situation is reversed: interpretations, recommendations, and actions are produced faster than humans can review them. The scarce resource is no longer insight, but the capacity to validate what the system produces — before an action is taken and while that action is still underway. This shift may be described as a transition from an economy of insight to an economy of validation, in which organisational performance increasingly depends on the reliability, controllability, and trustworthiness of decisions made with the assistance of artificial intelligence.

The consequences of this shift are most pronounced in high-risk and highly automated domains, such as mining, metallurgy, smart manufacturing, and financial systems. In such domains, an AI decision constitutes an action with immediate operational consequences. The speed and autonomy of these systems increase the risk of erroneous, biased, or unvalidated decisions propagating before anyone has time to respond. Under such conditions, validation is not a secondary control activity, but a core analytical function.

Avramović et al. [6] recently proposed the PRIME–INSPECT framework, a socio-technical model intended for governance, ethical alignment, and trust calibration in AI-supported decision-making in the context of Industry 4.0. This framework positions governance as the basic perspective from which AI-supported decisions are assessed. However, real-time validation is not

explicitly developed within it as an operational stage, nor is it situated within a closed decision loop. It is precisely this insufficiently developed aspect that forms the point of departure for the present paper, which directly builds upon the PRIME–INSPECT framework.

In response, this paper introduces an integrated conceptual framework for AI-native decision-making, composed of two parts. The first is the GAVA loop (Generate–Act–Validate–Adapt) — an operational, decision-theoretic model of the flow of AI-native processes, which runs continuously as the system generates decisions, executes actions, validates outcomes, and adapts to new information. The second part links the GAVA loop with the PRIME–INSPECT framework, thereby forming an architecture centred on validation and trust calibration, in which operational decision-making and socio-technical governance function as a single whole.

The contribution of this paper lies in three key respects. First, validation is positioned as the central organisational principle of AI-native systems, thereby shifting the analytical role of data science from the generation of insight to the performance of validation and control. Second, the operational perspective (GAVA) and the governance perspective (PRIME–INSPECT) are combined into a single framework that encompasses the interaction of algorithmic processes, human oversight, and institutional mechanisms. Finally, this integrated framework is empirically supported by inferential methods — multiple regression, mediation, moderation, and structural equation modelling — thereby continuing the future research programme outlined by Avramović et al. [6].

The paper is structured as follows, in accordance with the recommendations of Milutinović [7]. Section 2 presents the problem of real-time validation and the reasons why it is becoming increasingly important. Section 3 reviews existing solutions — explainable artificial intelligence and governance frameworks, reinforcement learning, control theory and cybernetics, and the PRIME–INSPECT framework — with a critical assessment of the limitations of each from the perspective of validation. Section 4 develops the proposed solution: the GAVA decision loop and its integration with the PRIME–INSPECT framework. Section 5 presents the empirical analysis. Section 6 discusses the theoretical, practical, and methodological implications and identifies limitations and directions for future research. Section 7 provides the conclusion, including a discussion of the implementation trade-offs of the architecture and the naming of the creative method applied.

2. Problem Definition

Real-time validation is becoming a core analytical function in AI-native decision systems.

The growing importance of real-time validation can be explained through three interrelated processes. The first is the increasingly rapid deployment of generative artificial intelligence and autonomous agents in operational environments: the volume and frequency of their outputs and decisions far exceed those of earlier rule-based or supervised-learning systems, while the share of decisions made without prior human review continues to increase. The second process arises from the application domains themselves — mining, metallurgy, smart manufacturing, financial trading, medical triage, and energy grids — which are subject to strict time constraints and leave little room for retrospective verification. The third is the increasingly stringent regulatory and ethical environment: the European Union Artificial Intelligence Act, sector-specific governance regimes, and stakeholder expectations require decisions that are explainable, accountable, and auditable under operational conditions.

The scale of this change is visible in practice. Research on the adoption of artificial intelligence in enterprises indicates a growing number of organisations using it not only for analytical support, but also for operational decision-making and task execution. At the same time, autonomous agents are increasingly moving from research prototypes into production environments, including customer support, content moderation, code generation, algorithmic trading, and clinical triage.

The pace of this change is particularly significant. The interval between the emergence of new generative capabilities and their practical deployment is becoming markedly shorter, leaving traditional cycles of verification, testing, and validation with ever less room to mature before AI-supported decisions begin to affect end users.

The regulatory environment is moving towards requirements that, in practical terms, mandate real-time validation. The European Union Artificial Intelligence Act, adopted in 2024, whose obligations for high-risk systems enter into force in 2026, requires continuous post-market monitoring of system performance, documented risk-management processes, and human oversight mechanisms throughout the entire life cycle. The U.S. NIST Artificial Intelligence Risk Management Framework (2023), together with the profile for generative artificial intelligence (2024), defines the functions of governing, mapping, measuring, and managing risk that must operate during system use, and not only at the design stage. Similar obligations are imposed by sector-specific

regulations — financial supervision of algorithmic trading, frameworks for medical devices with artificial-intelligence-based diagnostics, and occupational-safety requirements in autonomous manufacturing. Taken together, these regimes transform validation from an internal quality-control activity into an externally verifiable and continuously demonstrable property of the system, in line with recent calls for extreme AI risks to be regulated at the level of public policy [8].

The cost of an unvalidated decision increases with the degree of autonomy and speed. In supervised analytical environments, errors are most often detected by a human at the very point of decision-making; in AI-native systems, an incorrect decision is executed just as quickly as a correct one, before validation has even had time to occur. Recorded incidents confirm this: sudden market crashes in algorithmic trading, failures of automated content moderation, bias in automated credit approval, and erroneous triage in clinical systems [9], [10] have repeatedly concentrated harm within minutes rather than days. The gap between the speed of decision-making and the slowness of verification represents the operational core of this problem. For this reason, validation must be understood as a function embedded in the decision loop itself, rather than as a subsequent audit mechanism.

Under such conditions, validation cannot remain an ex post assessment activity. The interval between the emergence of a decision and its operational consequences is too short for verification outside the decision flow, while the cost of an unvalidated decision is potentially too high. Validation must therefore be continuous, embedded in the decision loop, and capable of assessing, in real time, the statistical accuracy, operational impact, cognitive intelligibility, and governance compliance of decisions.

The challenge can therefore be defined as follows. AI-native decision systems generate decisions at scale and execute them autonomously. Existing analytical paradigms are largely oriented either towards user acceptance of artificial intelligence or towards ex post model evaluation outside the actual decision flow, with neither of these approaches fully corresponding to the dynamics of real-time decision-making. What is missing is an integrated decision-theoretic framework in which validation has the status of a constitutive operational stage, embedded in a closed feedback loop and connected with mechanisms of governance and trust calibration. Designing such a framework is the central problem of this paper.

3. Existing Solutions

Four streams of prior research are relevant to the problem of validation presented in Section 2: explainable artificial intelligence and its governance, reinforcement learning, control theory and cybernetics, and the PRIME–INSPECT framework. Each of these captures one aspect of the problem, but none offers an integrated solution centred on validation. In this section, they are critically examined precisely from that perspective.

3.1 Explainable Artificial Intelligence and Governance

Research on explainable artificial intelligence (XAI), trust, and governance has highlighted important dimensions such as transparency, accountability, fairness, and risk management [11], [12], [13], [14]. Its core message is that an artificial intelligence system should be understandable, auditable, and aligned with ethical norms.

Viewed from a validation-centric perspective, however, this body of research still treats validation as a wrapper around a model developed independently. Explanations are, as a rule, generated after the model output, while governance is carried out through external audit rather than through mechanisms embedded in the operational decision loop itself. Validation is thus not formalised as a first-order operational stage, nor is it aligned with the speed and autonomy required by AI-native real-time systems.

3.2 Reinforcement Learning and Closed-Loop Systems

Reinforcement learning (RL) [15], [16] and related closed-loop paradigms are the closest theoretical antecedents of contemporary AI-native decision systems. An RL agent acts within a given environment and learns from feedback signals that shape its future behaviour. In this way, RL shares with these systems the basic logic of a closed structure, in which decisions, outcomes, and adaptation are connected through feedback.

From the standpoint of validation, the problem is that RL compresses feedback into a single scalar reward. Such an abstraction is extremely useful for learning, but it reduces the multidimensional nature of validation — statistical accuracy, operational indicators, cognitive intelligibility, and governance compliance — to a single number. In addition, RL generally operates through the relationship between an agent and an environment, while institutional constraints, regulatory obligations, and ethical alignment do not have a clearly defined place within the decision loop itself. More recent variants partially mitigate these limitations: constrained policy optimisation [17], safe reinforcement learning [18], and multi-

objective reinforcement learning [19] introduce safety constraints and multiple objectives into the learning process. Nevertheless, such approaches remain primarily within the bounds of decision theory and are not, in themselves, connected with broader institutional and governance mechanisms. In this sense, validation in RL is most often implicitly contained in the design of the reward function but is not explicitly separated as a distinct operational stage.

3.3 Control Theory and Cybernetics

Control theory [20] and the broader cybernetic tradition provide the canonical formalisation of feedback systems. A controller observes the state of the system, calculates a control signal, and applies it, after which the resulting state is measured and fed back in order to guide further control. Cybernetics generalises this picture to self-regulating systems that also adapt their own goals.

Here too, the validation-centric perspective reveals a limitation. Control theory and cybernetics formalise the loop, but leave the notion of feedback underspecified, reducing it to measurement and error correction. In a real-time AI-native system, the counterpart of error must include not only statistical conformity but also operational performance, human intelligibility, and institutional permissibility. Contemporary extensions mitigate these limitations to some extent. Model predictive control, with explicit state and input constraints [21], enables the formal encoding of operational and safety constraints. Similarly, supervisory architectures with human-in-the-loop adaptation [22] introduce additional mechanisms for oversight and correction of system behaviour. Even so, these formalisms continue to treat the controlled object primarily as a physical rather than a socio-technical system and do not directly express institutional, regulatory, or ethical conditions. The closed-loop formalism is therefore a necessary but not sufficient condition for validation in AI-native systems.

3.4 The PRIME–INSPECT Framework

Avramović et al. [6] introduced the PRIME–INSPECT framework as a socio-technical model of trustworthy intelligent automation in the context of Industry 4.0. Its PRIME component comprises the execution layer through the phases of prediction, regulation, interpretation, mitigation, and execution, while the INSPECT component comprises the governance layer through data integrity, navigability of explanations, supervisory control, policy and governance maturity, ethical compliance, collaboration, and trust calibration. The framework links algorithmic processes with organisational and institutional controls and was descriptively tested on a dual sample of IT professionals and top-management representatives. Full formal definitions, including the equations describing the PRIME execution flow and the INSPECT governance conditions, are provided in the original paper and are not repeated here.

From a validation-centric perspective, the PRIME–INSPECT framework establishes the institutional and ethical foundation necessary for trustworthy artificial intelligence, but does not formalise validation as an explicit first-order operational stage within a real-time decision loop. Within that framework, validation is implicitly contained in the Mitigate phase of the PRIME component and in the Supervisory Control component of the INSPECT framework. Significantly, Avramović et al. [6], in their section on future research, themselves emphasise the need to operationalise the dimensions of the PRIME–INSPECT framework on larger samples and to examine their interrelationships by means of inferential methods. The present paper develops the framework precisely in that direction: it embeds an explicit validation stage into a closed decision loop — the GAVA loop presented in Section 4 — and provides the inferential evidence that the original paper left for future research (Section 5).

3.5 MLOps and Operational Monitoring of Artificial Intelligence

The growing literature on MLOps and operational monitoring of artificial intelligence addresses the engineering procedures required for machine-learning systems to remain reliable in production. Sculley et al. [23] point to the hidden technical debt of such systems — the erosion of boundaries between components, implicitly defined consumers of system outputs, hidden feedback loops, and entangled data and model processing pipelines — while warning that maintenance costs may exceed the costs of initial development. Documentation tools, such as Model Cards [24], make visible a model's intended use, its limitations, and its performance disaggregated by subgroups. This body of research shows that production-grade artificial intelligence requires continuous monitoring of drift, performance degradation, and outcomes across subgroups. From a validation-centric perspective, MLOps already provides operational mechanisms for part of the functions that GAVA elevates to the level of an analytical stage; what it does not, by itself, provide is the integration of the cognitive and governance dimensions of validation, which the proposed framework treats as equal in status.

4. Proposed Solution: The Validation-Centric GAVA–PRIME–INSPECT Architecture

This section presents the proposed solution. First, the GAVA decision loop is introduced as an independent theoretical contribution in the field of decision-making, separate from the PRIME–INSPECT framework. The GAVA loop is then formalised in relation to reinforcement learning, control theory, and cybernetics, where it is presented as their generalisation in which validation is isolated as an explicit stage. After that, the GAVA loop is linked with the PRIME–INSPECT framework and the integrated validation-centric architecture is presented (Figure 1, Table 2). Finally, validation is defined as a constitutive analytical function based on four subdimensions.

4.1 The GAVA Decision Loop

GAVA is introduced as an integrative theoretical abstraction for understanding AI-native decision-making. The decision-making process is presented as a continuous four-phase cycle — Generate, Act, Validate, Adapt — unfolding in real time, with each phase conditioning and preparing the next.

Generate (G): The system produces a candidate decision based on a learned representation of the environment. The underlying mechanism may be a large language model, a predictive ensemble, or a hybrid processing pipeline; GAVA is agnostic in this respect and treats the model as a black box that provides both a decision and its rationale.

Act (A): The system executes the candidate decision in the environment, changing the state of the system and producing measurable outcomes.

Validate (V): The system evaluates the decision and its outcomes simultaneously along statistical, operational, cognitive, and governance dimensions, treating validation as a structured, multidimensional assessment rather than as a single scalar signal.

Adapt (A): On the basis of these multidimensional results, the system adapts both its own decision policy and its own validation criteria, thereby establishing the conditions for the next cycle.

Two features distinguish the GAVA loop from earlier closed-loop paradigms. The first is the multidimensionality of validation. Whereas reinforcement learning expresses feedback through a single reward and control theory through a single error signal, GAVA preserves four dimensions throughout the entire cycle — statistical accuracy, operational performance, cognitive intelligibility, and governance compliance — as distinct components. Collapsing them into a single number would discard precisely the information required for targeted adaptation, since different combinations of failure require different corrective actions. The second feature is that adaptation encompasses the validation criteria themselves, and not only the decision policy. Data drift, distributional shift, and regulation changes require the system to update not only the way it decides, but also the way it assesses whether a decision is good. In this sense, GAVA generalises reinforcement learning, control theory, and cybernetics: it retains their closed-loop structure, but enriches the feedback channel with structured, multidimensional information and allows validation criteria to evolve over time.

GAVA is therefore an abstraction, not a procedural recipe. It does not prescribe a specific model architecture, a specific validation method, or a specific adaptation rule. Its purpose is to provide a conceptual framework within which existing techniques — RL agents, explainable-AI tools, drift detectors, and governance checks — can be aligned into a coherent decision-making system in which validation is an integral operational stage.

4.2 Relationship with Reinforcement Learning, Control Theory, and Cybernetics

GAVA is consistent with established decision-theoretic paradigms and may be read as a generalisation of each of them. From the perspective of reinforcement learning, the Generate phase corresponds to policy generation, Act to action execution, Validate to the reward signal, and Adapt to policy updating. The first generalisation introduced by GAVA is the multidimensionality of the Validate phase instead of a scalar reward, while the second is the possibility that adaptation may modify the reward function itself.

From the perspective of control theory, Generate corresponds to the formulation of the control signal, Act to system actuation, Validate to measurement and error assessment, and Adapt to controller tuning. GAVA generalises classical control by treating measurement as a structured assessment along multiple axes and by allowing the controller's objective to change in accordance with governance constraints.

From a cybernetic perspective, GAVA realises a self-regulating system that observes its own outputs, compares them with a structured set of desired states, and adapts its behaviour accordingly. The four phases correspond to observation, comparison, decision-making, and behavioural adaptation in the classical cybernetic sense.

The relationship of the GAVA loop to the above-mentioned earlier closed-loop paradigms is summarised in Table 1.

Table 1. Mapping between Reinforcement Learning and the GAVA Decision Loop with Validation-Centric Extensions

Reinforcement Learning	GAVA	Extension provided by GAVA
Policy generation	Generate	AI reasoning (LLM-based or hybrid)
Action execution	Act	Autonomous execution under governance constraints
Reward signal (scalar)	Validate	Multi-dimensional validation: statistical, operational, cognitive, governance
Policy update	Adapt	Joint update of both the decision policy and the validation criteria

4.3 Integration with the PRIME–INSPECT Framework

GAVA describes the operational dynamics of decision-making, while PRIME–INSPECT captures the socio-technical governance that constrains and stabilises those dynamics. Their integration yields a validation-centric closed-loop architecture, shown in Figure 1.

Within this architecture, GAVA serves as the decision engine, while PRIME–INSPECT provides the execution and governance foundation. The Generate phase in the GAVA loop corresponds to the Predict component of PRIME, whose outputs are grounded in the INSPECT condition of Integrity of Data. The Act phase corresponds to the Execute component, constrained by the condition of Policy and Governance Maturity. The Validate phase corresponds to the Regulate component and relies on the conditions of Supervisory Control, Navigability, and Ethical Compliance — this is precisely where the multidimensional validation function V operates. The Adapt phase corresponds to the feedback loop of the PRIME framework and is stabilised by the conditions of Trust Calibration and Collaboration. This two-axis mapping is formalised in Table 2.

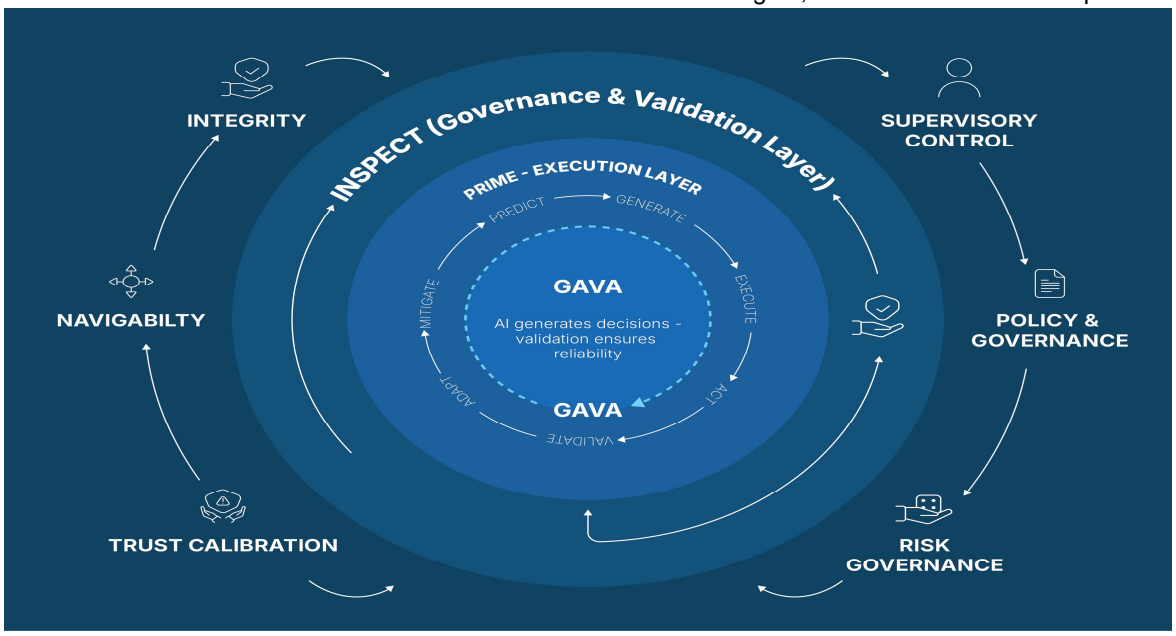

Figure 1. The integrated GAVA–PRIME–INSPECT validation-centric architecture. Legend: central GAVA loop; PRIME execution layer; INSPECT governance layer. Description: validation is embedded within the closed loop. Implication: autonomy and control achieved jointly through integrated operational/governance layers.

The integration is bidirectional, in accordance with the method of symbiosis [25]. From the PRIME–INSPECT framework, GAVA takes over the institutional and ethical foundation — supervisory control, trust calibration, ethical compliance, navigability, and collaboration — which the operational loop does not determine by itself. In return, from the GAVA loop, the PRIME–INSPECT framework receives an explicit operational closure of the validation feedback loop, which remained implicit in the original framework. Neither paradigm provides this capability on its own; only together do they determine what is validated and how the

institutional context evaluates and constrains the result. This also reveals a key structural property: in the GAVA–PRIME–INSPECT architecture, the closed loop is neither exclusively operational, as in reinforcement learning or control theory, nor exclusively governance-oriented, as in pure AI governance frameworks. It is both at the same time, with validation serving as the link between the two layers.

Table 2. Mapping between the GAVA Decision Loop and the PRIME–INSPECT Framework

GAVA Phase	PRIME Component	INSPECT Condition
Generate	Predict	Integrity of Data
Act	Execute	Policy and Governance
Validate	Regulate	Supervisory Control, Ethical Compliance
Adapt	Feedback loop	Trust Calibration

4.4 Validation as a First-Order Analytical Function

The integrated architecture elevates validation from an implicit, ex post activity to a central analytical function within the loop. Four subdimensions are distinguished — statistical, operational, cognitive, and governance validation — and the system evaluates them in every pass through the loop, with the validation function returning an assessment for each dimension and forwarding it to the Adapt phase. These four subdimensions may be illustrated by the example of a large-scale credit-scoring decision. Statistical validation checks the posterior calibration of the prediction and the range of uncertainty. Operational validation determines whether the decision keeps business indicators within prescribed boundaries — default rates, regulatory capital, and time budgets. Cognitive validation examines whether the explanation presented to the human in the loop is intelligible and usable for the specific operator, drawing on the literature on explainability and trust calibration [14], [26], [22]. Governance validation confirms that the decision is consistent with documented policies, ethical obligations, and audit trails [24]. Failure in any single dimension is sufficient to trigger the Adapt phase or, in the extreme case, a halt: the loop is not closed solely based on prediction error, but on the basis of the full multidimensional validation function.

This redefinition of validation also changes the analytical role of data science. In AI-native systems, the practitioner is no longer responsible only for the predictive model, but also for the validation infrastructure surrounding it — validation functions, drift detectors, explanation generators, governance mechanisms, and adaptation operators. This is precisely the shift described in Section 1 as the transition from an economy of insight to an economy of validation.

4.5 Research Hypotheses

The integrated GAVA–PRIME–INSPECT architecture implies a set of testable relationships among the socio-technical constructs that govern validation-centric decision-making. Drawing on the validation dimensions from Section 4.4 and on the constructs operationalised in the reused dataset of Avramović et al. [6], seven hypotheses are formulated and empirically tested in Section 5.

H1. Greater explainability of artificial intelligence (XAI) has a positive effect on trust in artificial intelligence.

H2. Greater explainability of artificial intelligence has a positive and direct effect on behavioural intention to adopt artificial intelligence in real-time decision-making.

H3. Trust in artificial intelligence mediates the relationship between its explainability and behavioural intention.

H4. Higher perceived risk has a negative effect on behavioural intention to adopt artificial intelligence in real-time decision-making.

H5. Greater top-management support has a positive effect on behavioural intention to adopt artificial intelligence in real-time decision-making.

H6. Artificial-intelligence literacy positively moderates the relationship between explainability and trust.

H7. Trust in artificial intelligence positively predicts perceived real-time decision-making performance (RTD), either directly or indirectly through behavioural intention. This relationship is treated as exploratory, since it is not estimated in the primary CB-SEM model.

Hypotheses H1–H3 operationalise the Validate phase of the GAVA loop and its cognitive validation subdimension, while hypotheses H4–H6 encompass the organisational conditions under which validation is maintained.

5. EMPIRICAL ANALYSIS

This section provides empirical support for the integrated GAVA–PRIME–INSPECT framework. It examines the relationships between explainability, trust, perceived risk, top-management support, behavioural intention, and real-time decision-making performance, presenting descriptive statistics, Pearson correlations, multiple linear regression, mediation and moderation tests, quadratic regression, group-comparison tests, and structural equation modelling.

5.1 Sample and Data

The analysis in this section is based on the

same dual sample previously presented in descriptive form by Avramović et al. [6]: 151 IT professionals and 85 top-management representatives. Whereas the original paper used this dataset to descriptively establish the recognisability of the PRIME–INSPECT framework dimensions as constructs, the present paper takes a further step by applying inferential techniques — multiple regression, mediation, moderation, and structural equation modelling — to test claims derived from the integrated GAVA–PRIME–INSPECT framework, in accordance with the direction for future research explicitly indicated in the original paper. Items from the original instrument were re-operationalised into seven constructs (XAI, Trust, Perceived Risk, AI Literacy, TMT Support, Behavioural Intention, RTD Performance) so as to reflect the validation-centric perspective. The XAI, Trust, and Behavioural Intention constructs were reconstructed from items previously grouped under broader operational and governance labels in [6], while the remaining constructs — Perceived Risk, AI Literacy, TMT Support, and RTD Performance — rely on partially overlapping sets of items with adapted labels. The two instruments (IT and TMT) differ in the number and wording of items per construct, reflecting the technical and strategic perspectives, respectively; this is acknowledged in the limitations and constitutes the reason for the robustness checks separated by subsample, presented below.

The data were collected from October to December 2025 through a structured online questionnaire distributed via professional networks and organisational contacts, using a combination of convenience and snowball sampling. Respondents came from organisations in which AI-native decision systems had already been introduced or were being actively considered.

The dataset consists of two respondent groups. The group of IT professionals (N≈151) comprises data scientists, machine-learning engineers, and IT managers, representing the technical-implementation perspective. The top-management team (TMT) group (N≈85) comprises senior executives responsible for strategic decision-making, governance, and oversight, representing the strategic perspective. All constructs were measured on a four-point Likert scale, ensuring group comparability and enabling direct comparative analysis. Participation was voluntary and anonymous.

This dual approach reflects the socio-technical nature of AI-native systems, in which technical capabilities and the organisational context jointly shape decision-making.

Table 3. Operationalisation of constructs and allocation of survey items, re-operationalised from [6].

Construct	**Representative indicators**	**IT instr.**	**TMT instr.**
XAI (explainability)	Understanding the reasons behind AI recommendations; verifying the factors that drive them	D1–D2	D1–D3
Trust	Confidence in AI outputs and their compliance with defined rules	D5–D6	D6, D10
Perceived Risk	Erroneous decisions, loss of control, data sensitivity, and ethical exposure	C1–C5	C1–C5
AI Literacy	Self-assessed AI/ML expertise (screening item)	Q3	Q4
TMT Support	Leadership support, accountability, and IT–business collaboration on AI	E1–E4	E1–E6
Behavioural Intention	Intention to expand AI use and embed it in new initiatives	B3–B5	B3–B5
RTD Performance	Perceived gains in decision speed, accuracy, efficiency, and ROI	H1–H4	G1–G5

Since the dataset originally presented by Avramović et al. [6] in descriptive form is re-operationalised here, Table 3 shows how each construct maps onto the corresponding items in both instruments. Explainability and trust were derived from items concerning intelligibility and confidence; perceived risk from the block concerning risk perception; top-management support from leadership-related items; behavioural intention from adoption-intention items; and RTD performance from items concerning perceived performance. AI literacy is captured through respondents' self-assessment of expertise in artificial intelligence and machine learning.

5.2 Descriptive Statistics and Reliability

Table 4 presents the means, standard deviations, and Cronbach's alpha coefficient for

each construct. All α values exceed the conventional threshold of 0.70 [27], confirming satisfactory internal consistency; the highest reliability is observed for the Trust construct (α = 0.88). The Trust, Behavioural Intention, and RTD Performance constructs have relatively high mean values, indicating a general openness towards AI-supported decision-making, while Perceived Risk shows the greatest variability, suggesting divergent attitudes towards risks associated with artificial intelligence.

Table 4. Descriptive statistics and reliability of measurement scales

Construct	Mean	SD	Cronbach's α
XAI	3.12	0.54	0.84
Trust	3.08	0.57	0.88
Perceived Risk	2.41	0.71	0.79
AI Literacy	2.95	0.60	0.82
TMT Support	2.87	0.65	0.85
Behavioural Intention (BI)	3.05	0.58	0.87
RTD Performance	3.01	0.55	0.86

5.3 Correlation Analysis

Table 5 presents the Pearson correlations among the seven constructs. The strongest positive relationship appears between Trust and RTD Performance ($r = 0.696$, $p < 0.01$), confirming the central role of trust in effective AI-supported decision-making. The strong correlation between XAI and Trust ($r = 0.645$) supports the role of explainability in building user trust [14]. Perceived Risk shows a significant negative correlation with Behavioural Intention ($r = -0.456$), while TMT Support correlates positively with the same construct ($r = 0.589$), underscoring the importance of organisational support. The sign of the correlations for the XAI construct differs from that of the corresponding aggregate in [6]: here, XAI has been re-operationalised around indicators of explainability and intelligibility — the perceived clarity of explanations — whereas in the earlier paper, related items were grouped into a broader operational-governance construct. This modification is consistent with the validation-centric perspective and is noted in the limitations.

Table 5. Pearson correlation matrix

Variable	1	2	3	4	5	6	7
1. XAI	1						
2. Trust	0.645**	1					
3. Risk	-0.312**	-0.421**	1				
4. AI Literacy	0.511**	0.533**	-0.274**	1			
5. TMT Support	0.402**	0.512**	-0.338**	0.441**	1		
6. BI	0.478**	0.601**	-0.456**	0.502**	0.589**	1	
7. Performance	0.522**	0.696**	-0.401**	0.488**	0.574**	0.612**	1

*Note: ** $p < 0.01$.*

5.4 Regression Analysis

Model 1 (Table 6) examines XAI as a predictor of Trust, controlling for AI Literacy. XAI has a strong and statistically significant effect on Trust ($\beta = 0.582$, $p < 0.001$), with $R^2 = 0.51$, thereby confirming the hypothesised XAI → Trust relationship.

Model 2 (Table 7) examines the determinants of Behavioural Intention. The strongest predictor is Trust ($\beta = 0.421$, $p < 0.001$). Perceived Risk acts in the opposite direction ($\beta = -0.287$, $p < 0.001$), in line with the findings of Glikson and Woolley [26], while TMT Support promotes it ($\beta = 0.318$, $p < 0.001$). The direct effect of XAI on BI is small ($\beta = 0.129$, $p = 0.028$), indicating that its influence operates predominantly indirectly, through trust.

Table 6. Regression results (DV: Trust)

Variable	β	SE	T	P
XAI	0.582	0.061	9.54	<0.001
AI Literacy	0.214	0.055	3.89	<0.001

$R^2 = 0.51$.

Table 7. Regression results (DV: Behavioural Intention)

Variable	β	SE	T	P
Trust	0.421	0.072	5.84	<0.001
Risk	-0.287	0.069	-4.16	<0.001
TMT Support	0.318	0.064	4.97	<0.001
XAI	0.129	0.058	2.22	0.028

$R^2 = 0.63$.

5.5 Mediation and Moderation

The mediation test (XAI → Trust → BI) yields an indirect effect of 0.102, with a 95% bootstrap

confidence interval of [0.021, 0.187] and $p < 0.01$. In combination with the modest direct effect of XAI on BI reported in Section 5.4, this indicates partial mediation: XAI affects behavioural intention primarily through trust.

The moderation test shows that AI Literacy strengthens the XAI → Trust relationship ($\beta = 0.167$, $p = 0.012$). Respondents with a higher level of artificial-intelligence literacy translate explainability into trust more effectively, which is consistent with the cognitive subdimension of validation introduced in Section 4.4.

As a robustness check of the linear Trust → Performance relationship, a quadratic specification including the Trust² term was examined. In the available data, no stable support was found for an inverted-U pattern of trust calibration. Trust calibration, therefore, remains a theoretically motivated construct that requires larger samples and longitudinal designs for confirmatory testing [22].

5.6 Comparative Analysis: IT and TMT

Welch's t-tests of group differences (Table 8) show that IT respondents report significantly higher trust than TMT respondents (3.21 versus 2.89, $t = 3.42$, $p < 0.001$). TMT respondents, on the other hand, perceive higher risk (2.58 versus 2.31, $t = -2.77$, $p = 0.006$) and attach greater importance to governance (3.02 versus 2.68, $t = -3.11$, $p = 0.002$). This pattern is consistent with the two-layer logic of the proposed architecture: the operational layer (GAVA) is closer to the perspective of IT professionals, while the governance layer (PRIME–INSPECT) reflects the priorities of top management.

Table 8. Group differences between IT and TMT respondents (Welch t-test)

Construct	IT Mean	TMT Mean	t	P
Trust	3.21	2.89	3.42	<0.001
Risk	2.31	2.58	-2.77	0.006
Governance	2.68	3.02	-3.11	0.002

5.7 Structural Equation Modelling

The structural equation model (SEM) estimates the direct and indirect relationships among the seven constructs simultaneously. The fit indices are $\chi^2(209) = 396.12$, $p < 0.001$, $\chi^2/df = 1.90$, CFI = 0.94, TLI = 0.93, RMSEA = 0.058, and SRMR = 0.047, exceeding the conventional acceptability threshold of 0.90 according to Hair et al. [27], although slightly below the stricter threshold of 0.95 proposed by Hu and Bentler [28]. To test possible measurement non-invariance between the two instruments, the same model was re-estimated on the IT subsample (N = 151), where it showed comparable fit (CFI = 0.94, TLI = 0.93, RMSEA = 0.060), while the TMT subsample (N = 85) showed acceptable but weaker fit (CFI = 0.85, RMSEA = 0.11), as expected given its smaller size. The structural paths reported below are stable across both specifications.

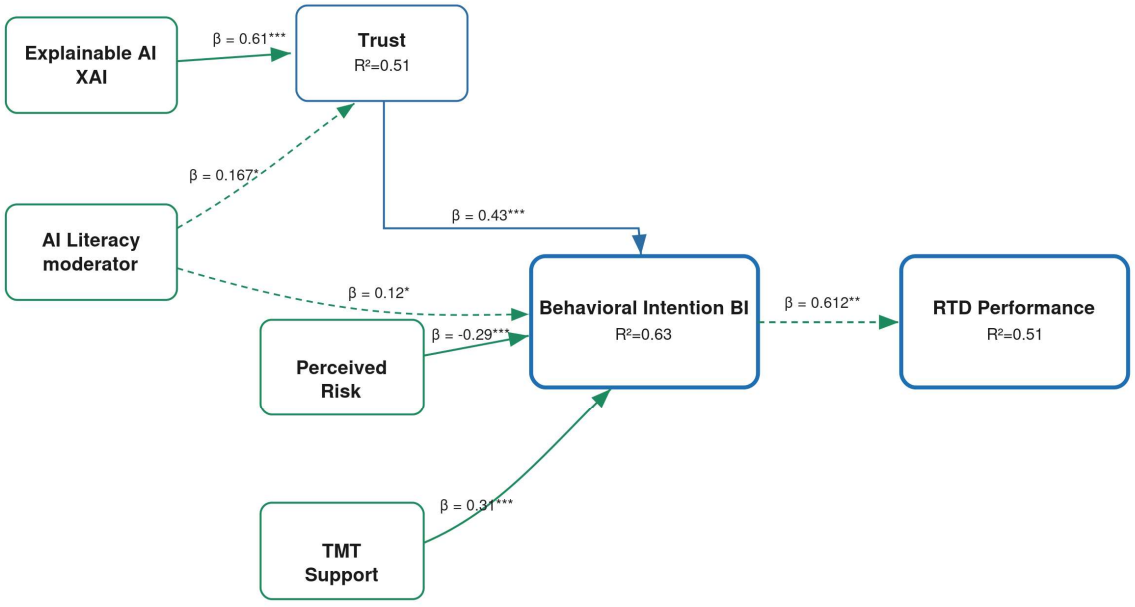

*Figure 2. Structural Equation Model and supplementary paths. Solid lines = paths estimated in the primary CB-SEM; dashed lines = moderation or supplementary correlation-based paths reported in Sections 5.4–5.5. Significance: *** p < 0.001, ** p < 0.01, * p < 0.05. Fit: CFI = 0.94, TLI = 0.93, RMSEA = 0.058, SRMR = 0.047.*

The SEM model confirms the findings obtained by regression: XAI → Trust ($\beta = 0.61$, $p < 0.001$); Trust → BI ($\beta = 0.43$, $p < 0.001$); Risk → BI ($\beta = -0.29$, $p < 0.001$); TMT Support → BI ($\beta = 0.31$, $p < 0.001$); XAI → BI ($\beta = 0.12$, $p = 0.031$). Figure 2 presents the full structural model with

standardised path coefficients.

The path from Behavioural Intention to RTD Performance (β = 0.612, p < 0.01) is presented as supplementary evidence, in accordance with the correlation reported in Section 5.3; it was estimated in a separate regression and is shown in Figure 2 as a dashed supplementary path, in line with the indicative support reported in [6]. The moderation effects of artificial-intelligence literacy (trust path β = 0.167, p = 0.012; supplementary BI path β = 0.12, p = 0.041) are also shown in Figure 2 as dashed lines and derive from the moderation specification presented in Section 5.5.

Although Cronbach's α for Trust in the TMT subsample was lower (α = 0.56), composite reliability weighted by standardised loadings reached the conventional threshold (CR = 0.82), so the construct remains acceptable for structural analysis despite the noise inherent in small samples.

Table 9. Composite reliability (CR) and average variance extracted (AVE) by subsample.

Construct	CR (IT)	AVE (IT)	CR (TMT)	AVE (TMT)
XAI	0.88	0.78	0.91	0.78
Trust	0.93	0.87	0.82	0.70
Perceived Risk	0.90	0.65	0.88	0.61
TMT Support	0.93	0.76	0.92	0.67
Behavioural Intention	0.91	0.77	0.97	0.91
RTD Performance	0.93	0.75	0.94	0.75

To supplement the reliability coefficients reported in Table 4, composite reliability (CR) and average variance extracted (AVE) were calculated for the six multi-item reflective constructs on the basis of the standardised measurement model (Table 9). All values exceed the recommended thresholds (CR > 0.70 and AVE > 0.50) [27], and standardised factor loadings range from 0.67 to 0.96, indicating satisfactory internal consistency and convergent validity in both subsamples. The AI Literacy construct, measured by a single self-assessment item, is treated as an observed control variable and is therefore omitted from the convergent-validity analysis. The weaker overall fit of the TMT subsample, noted above (N = 85; CFI = 0.85, RMSEA = 0.11), reflects the reduced stability of covariance-based estimation in small samples, rather than a deficiency of the measurement instrument, whose CR and AVE values remain satisfactory in both groups.

5.8 Summary of Empirical Findings

Four main findings emerge from the empirical analysis. First, explainability is a strong predictor of trust (β = 0.61). Second, trust mediates the effect of explainability on behavioural intention (indirect effect 0.102, CI [0.021, 0.187]). Third, perceived risk significantly slows adoption (β = -0.29). Fourth, top-management support significantly promotes it (β = 0.31).

Taken together, these results provide quantitative support for the central assumptions of the integrated GAVA–PRIME–INSPECT architecture. They show that validation-centric mechanisms — explainability, trust calibration, risk management, and governance more broadly — jointly shape the performance of AI-supported decision-making. At the same time, the results extend the findings of Avramović et al. [6], who reported only descriptive statistics on the same sample, by identifying the structural relationships assumed by the original framework.

These relationships map directly onto the GAVA loop and its validation subdimensions. Explainability (XAI) operationalises the Validate phase by providing intelligible evidence on which validation rests. Trust corresponds to the cognitive subdimension of validation and to the trust-calibration mechanism in the INSPECT layer. Perceived risk reflects governance validation and the regulatory boundaries that constrain autonomous action. Top-management support represents the organisational conditions that enable the Adapt phase and the supervisory function. RTD Performance captures the overall efficiency of the closed GAVA–PRIME–INSPECT loop. The empirical model therefore examines the socio-technical assumptions of validation-centric decision-making, rather than the lower-level operational parameters of the loop — such as validation latency or the dynamic adjustment of validation criteria — which remain the subject of future engineering evaluation.

The mapping of estimates onto the hypotheses is as follows: hypothesis H1 (XAI → Trust, β = 0.61) is strongly supported; hypothesis H3 is supported as partial mediation, with trust carrying the largest share of the effect of explainability on behavioural intention (Trust → BI, β = 0.43); the direct path in hypothesis H2 (XAI → BI, β = 0.12) is supported, but weakly, which is consistent with this mediation; hypotheses H4 (Risk → BI, β = −0.29) and H5 (TMT Support → BI, β = 0.31) are supported; and hypothesis H6 (AI Literacy moderates XAI → Trust, β = 0.167) is supported in the moderation specification from Section 5.5. The exploratory hypothesis H7 (Trust → RTD Performance) receives only indicative support through the strong correlation between trust and performance (r = 0.696) and is not estimated in the primary CB-SEM model, in accordance with [6].

6. DISCUSSION

6.1 Theoretical Implications

This paper contributes to the literature on AI-native systems by treating validation as an integral part of decision-making itself, rather than as a subsequent verification activity. Whereas earlier research has largely relied on insights derived from data and on model performance, the emphasis here shifts towards the question of whether AI outputs can be trusted while decisions are still being made and implemented.

The integration of the GAVA and PRIME–INSPECT frameworks moves the discussion in several important directions. The fundamental shift is conceptual in nature: validation is no longer treated as an external audit layer, but as part of the decision loop itself. In addition, this integration presents decision-making as a closed-loop process driven by feedback, thereby linking AI-native architectures with decision theory, reinforcement learning, and control systems, while placing governance, trust, and ethical constraints inside the loop rather than outside it.

The empirical findings support the socio-technical mechanisms on which this position rests, while the operational components of the closed loop remain the subject of future direct measurement. The significant effect of explainable artificial intelligence on trust ($\beta = 0.61$, $p < 0.001$) and the central role of trust in shaping behavioural intention ($\beta = 0.43$, $p < 0.001$) show that validation and intelligibility are not secondary attributes, but key drivers of artificial-intelligence adoption. The negative effect of perceived risk ($\beta = -0.29$, $p < 0.001$) and the positive influence of top-management support ($\beta = 0.31$, $p < 0.001$) further confirm the socio-technical character of the proposed framework.

In brief, the results show that effective AI-native systems require more than predictive capability: they require integrated mechanisms of validation, control, and trust calibration. This extends existing models of artificial-intelligence adoption and decision support.

6.2 Practical Implications

From a practical standpoint, the findings have direct consequences for how AI-native systems should be designed and introduced in organisations. Above all, organisations should not build such systems exclusively around prediction and automation. Accuracy remains important, but it is not sufficient when decisions are generated and executed in real time; it is equally important whether the system can explain, verify, and constrain its own outputs while they are being used.

This means that governance cannot be added only after the system has been introduced. Policies, ethical constraints, supervisory controls, and audit trails must be part of the system architecture from the outset.

The observed role of perceived risk and top-management support points to the same conclusion: the adoption of artificial intelligence depends not only on technical performance, but also on the organisation's ability to manage the risks introduced by automation.

Human oversight, too, should be understood dynamically and not as a final approval step. In validation-centric systems, human interpretation and trust calibration remain part of the loop, especially where decisions carry high-risk operational or organisational consequences.

For organisations, the message is clear: the scaling of AI-native decision systems requires investment in validation capabilities — performance monitoring, explainability tools, drift detection, and compliance checks — together with intervention mechanisms when validation fails. Without such investment, the scaling of systems may also mean the scaling of systemic risk.

6.3 Implications for Data Science

For data science, the implication is immediate. In AI-native environments, the work does not end at the moment when a model produces an accurate output; the more difficult task is to create the conditions under which that output can be monitored, explained, questioned, and corrected.

This shift is also confirmed by the empirically established role of trust. Trust mediates the relationship between explainability and behavioural intention, meaning that users respond not only to what the system recommends, but also to whether they can understand and assess that recommendation.

Data scientists are thereby moving increasingly closer to system governance. Their role increasingly includes validation design, real-time monitoring, bias and drift detection, ensuring explainability, supporting compliance, and defining escalation rules. In this sense, data science is becoming less a discipline of pure modelling and more a discipline of controlled decision-making.

6.4 Limitations

Since this paper combines the development of a conceptual framework with empirical analysis, it is subject to several limitations.

First, although the results support the socio-technical mechanisms of the proposed framework, they are based on cross-sectional data, which limits the possibility of drawing conclusions about causal relationships over time.

In addition, the integration of the GAVA and PRIME–INSPECT frameworks is presented at a relatively high level of abstraction. Such an approach is conducive to general applicability, but may not fully capture domain-specific constraints — regulatory requirements, technical architectures, or risk factors characteristic of particular industries.

The sample, although diverse, is limited to two respondent groups (IT and TMT) and may not represent all stakeholders involved in AI-assisted decision-making processes.

Finally, the implementation of validation-centric systems depends on organisational maturity, data quality, and infrastructural capabilities, which may vary considerably across contexts.

6.5 Future Research

The proposed framework opens several directions for future research.

First, longitudinal studies are needed to examine how mechanisms of trust and validation develop in AI-native systems over time.

Second, trust calibration deserves deeper investigation, particularly with respect to longitudinal dynamics and the dynamics of overreliance, which a cross-sectional design cannot capture.

Further, it would be worthwhile to examine applications of the framework in specific domains, especially in high-risk environments such as manufacturing, finance, healthcare, and energy.

Additional work is also needed on the operationalisation of validation-centric architectures, including the development of metrics, tools, and guidelines for implementation in real-time decision-making environments.

Finally, the changing role of data science in AI-native systems requires closer consideration, particularly with regard to organisational structures, required skills, and interdisciplinary collaboration.

7. Conclusion

This paper examined decision-making in AI-native systems, in which artificial intelligence increasingly not only supports human judgement but also independently generates and executes decisions. In response to this shift, a validation-centric framework was proposed, linking the GAVA decision loop with the socio-technical PRIME–INSPECT model [6].

The central argument of the paper is that validation should be regarded as an integral part of the decision-making process itself. In traditional data-driven environments, the main challenge was most often to arrive at useful insights; in AI-native systems, the more difficult question is whether the decisions they generate can be verified, constrained, and trusted while action is being taken upon them. Embedding validation in a closed decision loop and linking it with governance represents one way of addressing this problem.

The integration of the GAVA and PRIME–INSPECT frameworks is useful precisely because the two perspectives cover different parts of the same challenge. GAVA describes the operational cycle of generation, action, validation, and adaptation, while PRIME–INSPECT adds the governance, trust, and oversight conditions that make that cycle accountable. The empirical results support this interpretation, demonstrating the role of explainability, trust, perceived risk, and top-management support in AI-supported decision-making.

The proposed architecture also comes at a cost. It requires additional infrastructure — for multidimensional validation and for the adaptation of both decision policies and validation criteria — in exchange for gains in reliability, intelligibility, and governance compliance. In high-risk domains, such as mining, finance, smart manufacturing, and healthcare, this trade-off will most often be worthwhile.

In methodological terms, the contribution of the paper can be understood through the method of symbiosis [25]: two previously separate paradigms — operational decision loops and socio-technical governance — are connected to achieve a new capability, real-time validation as a first-order analytical function, which neither of them provides on its own. A secondary method is interdisciplinary application, that is, the transfer of abstractions from reinforcement learning and control theory into the field of artificial-intelligence governance and organisational decision-making.

Taken as a whole, the paper offers a validation-centric perspective on AI-native systems, aligning algorithmic decision-making with organisational and societal requirements. As these systems grow in complexity and autonomy, future research should further focus on empirical validation across different domains, the dynamics of trust calibration, and the development of operational validation frameworks for real-time decision-making environments.

Declaration on the Use of Generative Artificial Intelligence

The authors confirm that generative artificial-intelligence tools were used solely to improve the written expression, readability, and linguistic fluency of the text. Their use was limited to sentence reformulation and to improving clarity and coherence. All content produced by these tools was carefully reviewed, edited, and

integrated by the authors, who assume full responsibility for the content of the manuscript.

REFERENCES

[1] Bommasani, R., Hudson, D. A., Adeli, E., et al., "On the opportunities and risks of foundation models," Stanford Center for Research on Foundation Models (CRFM), arXiv:2108.07258, 2021, doi: 10.48550/arXiv.2108.07258.

[2] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y., "ReAct: Synergizing reasoning and acting in language models," in International Conference on Learning Representations (ICLR), 2023, arXiv:2210.03629.

[3] Duan, Y., Edwards, J. S., and Dwivedi, Y. K., "Artificial intelligence for decision making in the era of big data — Evolution, challenges and research agenda," International Journal of Information Management, vol. 48, pp. 63–71, 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[4] Dwivedi, Y. K., Hughes, L., Ismagilova, E., et al., "Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," International Journal of Information Management, vol. 57, art. no. 101994, 2021, doi: 10.1016/j.ijinfomgt.2019.08.002.

[5] Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach, 4th ed., global ed. Harlow, U.K.: Pearson, 2021.

[6] Avramović, N., Marković, A., Čomić, T., Čavoški, S., Zornić, N., and Vujović, V., "PRIME–INSPECT: A Socio-Technical Framework for Trustworthy Intelligent Automation and Real-Time Decision-Making in Industry 4.0," Applied Sciences, vol. 16, no. 10, art. no. 4825, 2026, doi: 10.3390/app16104825.

[7] Milutinović, V., "The best method for presentation of research results," IEEE TCCA Newsletter, no. 9, pp. 1–6, 1996.

[8] Bengio, Y., Hinton, G., Yao, A., et al., "Managing extreme AI risks amid rapid progress," Science, vol. 384, no. 6698, pp. 842–845, 2024, doi: 10.1126/science.adn0117.

[9] Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T., "The flash crash: High-frequency trading in an electronic market," Journal of Finance, vol. 72, no. 3, pp. 967–998, 2017, doi: 10.1111/jofi.12498.

[10] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S., "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.

[11] Floridi, L. and Cowls, J., "A unified framework of five principles for AI in society," Harvard Data Science Review, vol. 1, no. 1, 2019, doi: 10.1162/99608f92.8cd550d1.

[12] Giudici, P., Centurelli, M., and Turchetta, S., "Artificial intelligence risk measurement," Expert Systems with Applications, vol. 235, art. no. 121220, 2024, doi: 10.1016/j.eswa.2023.121220.

[13] Selbst, A. D. and Barocas, S., "The intuitive appeal of explainable machines," Fordham Law Review, vol. 87, no. 3, pp. 1085–1139, 2018.

[14] Shin, D., "User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability," Journal of Broadcasting & Electronic Media, vol. 64, no. 4, pp. 541–565, 2020, doi: 10.1080/08838151.2020.1843357.

[15] Mnih, V., Kavukcuoglu, K., Silver, D., et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, pp. 529–533, 2015, doi: 10.1038/nature14236.

[16] Sutton, R. S. and Barto, A. G., Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[17] Achiam, J., Held, D., Tamar, A., and Abbeel, P., "Constrained policy optimization," in Proceedings of the 34th International Conference on Machine Learning (ICML), vol. 70, 2017, pp. 22–31, arXiv:1705.10528.

[18] García, J. and Fernández, F., "A comprehensive survey on safe reinforcement learning," Journal of Machine Learning Research, vol. 16, pp. 1437–1480, 2015.

[19] Hayes, C. F., Rădulescu, R., Bargiacchi, E., et al., "A practical guide to multi-objective reinforcement learning and planning," Autonomous Agents and Multi-Agent Systems, vol. 36, no. 1, art. no. 26, 2022, doi: 10.1007/s10458-022-09552-y.

[20] Åström, K. J. and Murray, R. M., Feedback Systems: An Introduction for Scientists and Engineers, 2nd ed. Princeton, NJ, USA: Princeton University Press, 2021.

[21] Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E. N., "The explicit linear quadratic regulator for constrained systems," Automatica, vol. 38, no. 1, pp. 3–20, 2002, doi: 10.1016/S0005-1098(01)00174-1.

[22] Lee, J. D. and See, K. A., "Trust in automation: Designing for appropriate reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50.30392.

[23] Sculley, D., Holt, G., Golovin, D., et al., "Hidden technical debt in machine learning systems," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015, pp. 2503–2511.

[24] Mitchell, M., Wu, S., Zaldivar, A., et al., "Model cards for model reporting," in Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*), 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[25] Blagojević, V., Bojić, D., Bojović, M., Cvetanović, M., Đorđević, J., Đurđević, Đ., et al., "A systematic approach to generation of new ideas for PhD research in computing," in Advances in Computers, vol. 104, A. R. Hurson and V. Milutinović, Eds. Cambridge, MA, USA: Academic Press, 2017, pp. 1–31.

[26] Glikson, E. and Woolley, A. W., "Human trust in artificial intelligence: Review of empirical research," Academy of Management Annals, vol. 14, no. 2, pp. 627–660, 2020, doi: 10.5465/annals.2018.0057.

[27] Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E., Multivariate Data Analysis, 8th ed. Andover, U.K.: Cengage Learning, 2019.

[28] Hu, L. and Bentler, P. M., "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Structural Equation Modeling, vol. 6, no. 1, pp. 1–55, 1999, doi: 10.1080/10705519909540118.

AUTHOR BIOGRAPHIES

Nebojša Avramović

Nebojša Avramović is the founder and CEO of Consulting IT DNA, a company specialised in business and digital transformations through the application of artificial intelligence and modern technologies. He graduated from the Military Technical Academy in Belgrade and is currently a doctoral candidate at the Faculty of Organisational Sciences, University of Belgrade, in the Information Systems and Quantitative Management programme. Throughout his career he has held management positions at Oracle, Ingram Micro, and IBM. He has co-authored papers on agent-based models, artificial intelligence, and digital technologies in business systems. E-mail: nebojsa.avramovic@consultingitdna.rs, ORCID: 0009-0009-2565-7176.

Tijana Čomić

Tijana Čomić, PhD, is an Assistant Professor and researcher specializing in statistics, quantitative analysis, survey methodology, and official statistics. She earned her PhD from the University of Belgrade, Faculty of Organisational Sciences, with a dissertation on improving official statistics through Big Data applications. Her academic work includes publications in international journals and conference proceedings, with research interests spanning sustainable development indicators, survey methodology, disability statistics, artificial intelligence, and data-driven decision-making. She has extensive experience in teaching quantitative subjects and applying advanced statistical methods in international research and development projects. E-mail: tijana.comic@vos.edu.rs, ORCID: 0000-0001-6556-5879.

Aleksandar Marković

Aleksandar Marković is a full professor at the Faculty of Organisational Sciences, University of Belgrade. He teaches Simulation in Business Decision Making, Financial Simulation Modeling, Business Dynamics, and E-business Management. He served as Vice-Dean for Organisation and Finance (2016–2021) and Editor-in-Chief of the international journal Management (2007–2016). His research focuses on business systems modelling, computer simulation, system dynamics, and e-business. He has published more than 70 scientific papers. E-mail: aleksandar.markovic@fon.bg.ac.rs, ORCID: 0000-0002-3976-2788.

Sava Čavoški

Sava Čavoški is an Assistant Professor at the School of Computing, Union University. His professional and research interests focus on computer simulation, agent-based modelling and simulation, machine learning, and AI-supported decision-support tools. He has experience in the design and development of complex enterprise software systems, scalable information systems, and cloud-based applications. His current work focuses on generative agent-based modelling (GABM), the strategic development of information systems, and the integration of artificial intelligence into business decision-making and digital transformation processes. E-mail: scavoski@raf.rs, ORCID: 0000-0003-1163-4466.

Nikola Zornić

Nikola Zornić, PhD, is an Assistant Professor at the Faculty of Organisational Sciences, University of Belgrade, where he received his bachelor's and master's degrees and defended his doctoral dissertation in 2023 on a methodological framework for integrating machine learning algorithms into agent-based simulation models. His research interests span agent-based modelling and simulation, machine learning, scientometrics, and business analytics, with current work focused on generative agent-based modelling (GABM). He has participated in several European and national research projects and has served on the Faculty Council since 2022. E-mail: nikola.zornic@fon.bg.ac.rs, ORCID: 0000-0002-3597-0627.

Vladimir Vujović

Vladimir Vujović is an assistant professor at the Faculty of Electrical Engineering, University of East Sarajevo, in the field of Computer Science. He has authored more than 80 scientific and professional papers, participated in several research projects, and is an active member of professional, scientific, and conference committees. His professional interests focus on software architecture, business analysis, and strategic development of information. E-mail: systems.vladimir.vujovic@etf.ues.rs.ba, ORCID: 0000-0001-8175-5058.