

HSE UNIVERSITY

**THE INTERNATIONAL LABORATORY FOR DIGITAL
TRANSFORMATION IN PUBLIC ADMINISTRATION**

Conference Papers of the 3rd

Digital Governance and Public Administration:
Global and Interdisciplinary Perspectives conference

June 18-19th, 2026

“Intelligent Governance: Transformations in the Digital Age”

Conference Organizers:

Eran Vigoda-Gadot – The International Laboratory for Digital Transformation in Public Administration, HSE University Academic Supervisor; Professor of the Department of Public Administration and Politics at the School of Political Sciences of the University of Haifa. Israel.

Evgeny Styrin – Head of the International Laboratory for Digital Transformation in Public Administration, HSE University, Moscow, Russia.

Conference Papers were edited by **Maria Yudina**, Senior Research Fellow of the International Laboratory for Digital Transformation in Public Administration, HSE University, Moscow, Russia

Table of contents

18 JUNE 2026 SESSION 1	2
Cross-Border Data Flows and AI Cooperation between China and Russia	2
Leaping the Digital Chasm: Interdisciplinary Governance, Regional Interoperability, and the Southeast Asian GovTech Renaissance	2
Performance Feedback and Strategic AI Governance Framing	4
Studying Digital Governance beyond Traditional Public Administration: Emotional, Platform-Based and Methodological Challenges.....	5
18 JUNE 2026 SESSION 2	6
Digital Transformation in Brazil: Federative and Cross-Sectoral Coordination... 6	
(Social) Fintechs, Digital Governance, and the Regulatory Challenges of Financial Inclusion	7
Agentic Governance in Local Public Administration: The Experience of the Municipality of Pedro Leopoldo, Brazil	7
19 JUNE 2026 SESSION 1	8
Both User and Citizen: Data Protection and Data Collaboration in the Co-Production of Public Services	8
Digital Platforms in Public Administration: Enabling Intelligent Governance through Data Integration and Institutional Transformation.....	9
Integral Assessment of the Effectiveness of Industrial Goods Promotion as a Tool for Intelligent Management in the Digital Economy	10
The Image of Authority in Digital Format: How Citizens Perceive the Names of Chatbots for Public Services?	11
Artificial intelligence in public administration: balancing efficiency and ethical norms.....	11
19 JUNE 2026 SESSION 2	12
The Digital State and the Fragmentation of the Global Digital Economy: Institutional Models and Digital Sovereignty	12
The Problem of Trust in Legal Artificial Intelligence Solutions in the Context of Ensuring Human and Civil Rights and Freedoms	13
Tools for Protecting Against Reward Function Hacking and Systemic Corruption in Artificial Intelligence of Government Systems	15
When an Algorithm Evaluates People: AI-Assisted Management and the Limits of Algorithmic Assessment.....	16

18 JUNE 2026 SESSION 1

Cross-Border Data Flows and AI Cooperation between China and Russia

Prof. Tao Chen, School of Public Administration, Huazhong University of Science and Technology, Wuhan, China

The global digital economy runs on data. Understanding the scale, speed and value of cross-border data flows sets the stage for everything that follows. Average annual growth rate of digital trade (2020-2024) is 12.1%, and the share of global digital trade is 21.9% in total international trade 2024. China, the EU, and US form the core hub of global digital trade, while emerging economies (India, Brazil, Mexico) are growing at 14-17% annually, reshaping the global landscape.

Data as new oil create value and growth. Smooth data flows cut through 'iceberg costs' of trade – the hidden friction that slows transactions. Scale and network effects boost productivity for multinational businesses. China sits at the heart of global digital services value chains. Digital platforms have lowered the barrier to entry, letting small and medium businesses sell directly to the world.

China risk-based governance helped to create 2024 breakthrough. The 2024 rules introduced clear thresholds. For example, if a company handles fewer than 10,000 people's personnel data or fewer than 1,000 people's sensitive data, they may be exempt from heavy security assessments. This removes red tape for routine HR, logistics, and contact work.

Leaping the Digital Chasm: Interdisciplinary Governance, Regional Interoperability, and the Southeast Asian GovTech Renaissance

Prof. Deden Witarsyah, Universiti Tun Hussein Onn Malaysia and Nusa Putra

The realization of Digital-Era Governance is inherently complicated by a phenomenon identified in academic literature as the "Digital Governance Paradox". The paradox dictates that as public administration becomes increasingly automated, data-driven, and reliant on third-party digital platforms, the technological distance between the state and the citizen shrinks, generating unprecedented levels of administrative efficiency. Yet, sociologically, this exact same process has the profound potential to widen the distance between the state and the citizen if digital trust is broken, if privacy is compromised, if algorithmic bias goes unchecked, or if digital equity is ignored. The paradox highlights a chilling reality: technologies that are explicitly positioned as facilitators of civic participation, governmental transparency, and streamlined service delivery can simultaneously function as sophisticated tools for digital surveillance, state domination, and administrative centralization.

Public administrative bodies face three core friction points when designing systems that attempt to resolve the paradox "Digital Governance Paradox". The tension between technical ambition and governance reality dictates the success or failure of digital statecraft. Synthesizing the exhaustive empirical evidence, theoretical frameworks, and institutional case studies drawn from the Southeast Asian GovTech renaissance, a clear, highly actionable framework emerges for the global public administration community. As sovereign states architect governance models toward the 2030 milestones, the following interdisciplinary strategies must be prioritized to ensure resilience, efficiency, and equity:

- **Shift from "Siloed IT" to "Whole-of-Government" Platforms:** The era of funding isolated, proprietary software solutions for individual ministries is obsolete and fiscally irresponsible. Public value is exponentially maximized when states construct foundational, open, and shared Digital Public Infrastructure (DPI). Centralized layers for unified Digital Identity, Data Exchange, and Payments – as demonstrated by Indonesia's INA Digital consolidation and Singapore's highly integrated Singpass – must serve as the interoperable backbone of the state, drastically reducing technical redundancy and citizen friction.

- **Institutionalize Agile Policy Making and Sandboxes:** Technology evolves at an exponential curve, while traditional bureaucracy moves linearly. To prevent regulatory frameworks from instantly becoming archaic, public agencies must aggressively adopt anticipatory governance models. The deployment of regulatory sandboxes – controlled environments to test policies prior to formal codification, as utilized in Vietnam's AI law rollout and Singapore's Trust Tech initiatives – allows the state to foster rapid innovation while meticulously managing public risk and ensuring compliance.

- **Prioritize the Development of Hybrid Public Servants (Civil Servant 2.0):** The modern state requires a fundamental paradigm shift in human capital. There is an acute global necessity for a new generation of bureaucrats who are bilingual in both administrative public policy and technical data architecture. Policymakers must intimately understand the mechanics of algorithmic bias, cryptographic security, and database interoperability to effectively govern the Agentic State, execute business process reengineering (BPR), and engage in meaningful, secure technology procurement.

- **Enshrine Human-Centric Co-Production:** The ultimate success and legitimacy of digital governance rest entirely on continuous, meaningful civic engagement. Citizens must be viewed as active partners and co-producers of public value, not passive recipients or data points. Digital platforms must incorporate direct, iterative feedback loops and prioritize universal accessibility. By actively co-designing services with the public, governments ensure that digital interventions serve the most vulnerable and marginalized populations, thereby mitigating the digital paradox and preventing technology from exacerbating existing socioeconomic divides.

The future trajectory of public administration is actively being written in the code, protocols, and policies of the digital era. It is an evolution that is inherently data-driven, highly interdisciplinary, and increasingly borderless. The rapid advancement of digital governance in Southeast Asia clearly and empirically demonstrates that immense structural challenges,

historical bureaucratic rigidities, and severe developmental divides can be transformed into profound opportunities for rapid innovation and leapfrogging. Through landmark, paradigm-shifting regional agreements like the ASEAN Digital Economy Framework Agreement (DEFA), and bold, aggressive domestic reforms ranging from Indonesia's GovTech consolidation (INA Digital) to Vietnam's pioneering AI accountability frameworks, the ASEAN region is actively defining the operational parameters of the modern digital state.

However, technological deployment without corresponding, deep-rooted institutional and cultural reform is entirely insufficient. As evidenced by the persistent digital governance paradox, states must meticulously and continuously balance the pursuit of raw administrative efficiency with the immutable rights of citizen privacy, algorithmic transparency, and broad social equity. By wholeheartedly embracing the advanced principles of Digital-Era Governance and the active co-production of public value, governments globally can architect digital states that are not merely faster and more cost-effective, but are fundamentally fairer, highly resilient, and profoundly human-centric.

Performance Feedback and Strategic AI Governance Framing

Ziteng Fan, Associate Professor, Fudan University, Shanghai, China

While AI is expected to transform government performance, less is known about how governments' existing performance information shape the ways they interpret, frame, and respond to AI as an uncertain policy issue. The governance of emerging technologies often begins with how stakeholders frame them. Issue framing captures the way by which governments interpret a certain policy issues, representing an important way to respond to uncertainty. Governments increasingly face the task of governing emerging technologies whose social and administrative meanings are not yet settled. Governments can frame AI in different ways. This leads to a dichotomy: Opportunity-oriented v.s. Risk-oriented framing of AI governance. These different framings matter because they shape what governments attend to, what policy problems they recognize, and what kinds of governance responses become legitimate.

Performance management is now regarded as one of the most important practices. Governments can use performance information to enable informed decision-making, assess achievements, set targets, and make strategic adjustments. How do governments use performance feedback to shape their issue framing strategies in AI governance? Existing studies generally assume that governments respond to performance information by changing what they do: reallocating resources, revising policies, or shifting managerial priorities (Greve 2003; Hong et al. 2019; Kim 2023; Huang 2024; Chamber et al. 2025). Less attention has been paid to how

performance feedback shapes how governments make sense of uncertain policy issues before concrete policy action takes place.

In innovation-driven domains such as digital transformation, governments are more responsive to positive performance feedback, rather than negative feedback. Quantitative positive feedback may have a negative effect on the prevalence of defending frames in AI discourse. Exemplary positive feedback, such as public recognition of best practices, may positively influence the prevalence of defending frames in AI discourse.

Studying Digital Governance beyond Traditional Public Administration: Emotional, Platform-Based and Methodological Challenges

Laboratory Head Evgeny Styrin; Associate Professor, Leading Research Fellow Anna Sanina, International Laboratory for Digital Transformation in Public Administration, HSE University, Moscow, Russia

Digital Governance can be called a competence challenge, because it emerges at the intersection of multiple disciplines. Research questions increasingly require new methods and forms of evidence. The challenge is not only what we study, but how we study it. So, digital governance is a methodologically hybrid field. The table below shows how it affects studying public administration in modern days.

Public administration question	In digital governance, the questions are	Methods and fields involved
Trust	How do citizens trust AI-enabled public services?	AI, data science, behavioral experiments
Accountability	Who is responsible for algorithmic decisions?	Information systems, algorithm audits, platform governance
Public value	Do digital services create value for citizens, not only efficiency?	Platform studies, service design, digital trace data
Legitimacy	Are automated and data-driven decisions seen as fair and acceptable?	Psychology, behavioral science, sentiment analysis
Citizen-state relations	How do citizens experience interaction with the digital state?	UX / HCI, user research, digital interaction data

This diversity of questions and skills required to study those leads to the competence boundary problem. It can be formulated in question for each level:

Individual Level: What should a digital governance scholar master?

Collaboration Level: What requires interdisciplinary teamwork?

International Level: Do competence expectations differ across countries and academic traditions?

Therefore digital governance challenges not only methods, but also the organization of expertise. It leads to research agenda with the key question: how are competence boundaries

defined across different academic communities? The International Laboratory for Digital Transformation in Public Administration, HSE University, is open for collaboration on the matter.

18 JUNE 2026 SESSION 2

Digital Transformation in Brazil: Federative and Cross-Sectoral Coordination

Prof. Lizandro Lui, School of Public Policy and Government, FGV University, Brazil;

Denise Direito Ph.D., Specialist in Public Policy and Government Management, IPEA, Brazil

Technology favors centralization: digital systems benefit from massive data aggregation, as AI algorithms, unified authentication, interoperability, and predictive analytics achieve better outcomes when fed by large volumes of information. Brazilian federalism grants autonomy to states and municipalities creates, which creates a challenge in balancing coordination and autonomy. One of these troubles aspect is the horizontal coordination. For example, Federative Strategies for Digital Government (EFGC): it is fine normatively, but sectoral policies like Health, Education, Social Assistance, are consolidated areas with a tradition of autonomous agreements and execution. There is a difficulty in adopting common interoperability standards, and culture of silos.

"State as a Platform" model reimagines government as a set of shared digital utilities that private and public actors can build upon, reducing duplication and improving user experience. In Brazil the federal government adheres pro forma: it digitizes existing analog services without rethinking or transforming them. Digital Civil Identity is still under construction. Began in 1997 it faced the veto power: states control civil ID issuance. Only 22 million National Identity Card (CIN) were issued which is approximately 10% of population. Gov.br authenticator ecosystem has 176 million Gov.br accounts. Only 35% of municipalities have adhered to the Gov.br. Those examples show how Brazil National Digital Government Strategy faces the Federative Challenge

The advancement of Gov.br has consolidated important capabilities at the federal level. However, Brazil is a federation in which public policies and service delivery depend on the joint action of the federal government, states, and municipalities. Digital maturity remains highly uneven: while some subnational governments have developed sophisticated digital solutions, others face significant technical, financial, and organizational constraints. A coordination strategy based primarily on voluntary adherence, the dissemination of best practices, and soft incentives has produced heterogeneous outcomes. While states such as Rio Grande do Sul, Goiás, and São Paulo have developed more advanced and integrated digital ecosystems, others states continue to struggle with obsolete structures and limited technical capacity.

(Social) Fintechs, Digital Governance, and the Regulatory Challenges of Financial Inclusion

Prof. Erica Siqueira, Associate Professor, Fundação Getulio Vargas — FGV EAESP, São Paulo, Brazil FGV

These theses discuss the regulatory and governance challenges associated with digital financial inclusion, focusing on fintech companies that combine financial technologies, social impact discourse, and for-profit business models. The social discourse of fintech companies claims that digital finance can promote inclusion, sustainability, and socioeconomic development. However this claim should be analyzed with caution, because there is also “surveilled inclusion.” Based on research published in the Journal of the Association for Information Systems, I argue that digital financial inclusion may expand access while simultaneously intensifying data extraction, behavioral monitoring, and new forms of dependency among low-income populations. Social fintechs combine explicit social missions with market-oriented, data-driven, venture-capital-backed business models. Extended surveillance – digital, community, and personal – is structural to this business model, expand platform control over borrowers: digital surveillance, community surveillance, personal surveillance. Low-value, short-term microcredit is designed to pressure renewal, generating continuous cycles of payment and borrowing that serve platform growth metrics (LTV, CAC).

The distance between social mission discourse and operational effects is structural – grounded in how these organizations are financed and evaluated. Financial inclusion, as operationalized by these platforms, serves as the discursive frame for a business model built on continuous cycles of debt. I address the relationship between the state, entrepreneurship, and financialized inclusion drawing on research published in Environment and Planning A: Economy and Space. It examines how digital and financial inclusion agendas may transfer responsibility for survival to individuals, especially informal workers and populations living in contexts of socioeconomic vulnerability. An ongoing research project on loose regulatory frameworks designed to foster financial inclusion and competition. Preliminary findings suggest that, while such frameworks may encourage innovation and market entry, they can also create conditions for fraud, scams, and new risks for vulnerable consumers.

Agentic Governance in Local Public Administration: The Experience of the Municipality of Pedro Leopoldo, Brazil

Dr. Daniel de Souza Valotto, Ms. Lorryne Silva, Specialist. Municipality of Pedro Leopoldo, Brazil

Brazil is highlighted by Digital Maturity in Rankings, however there is a digital and capability gap between major capitals and small/medium-sized municipalities. Fiscal constraints, fragmented legacy systems, and shortage of specialized IT personnel are problems State of Minas Gerais (Southeast Brazil) faces. State of Minas Gerais economy focus on Industry (Textiles & Cement) and Agribusiness, and identity are connected with "Luzia," the oldest human fossil found in South America (12,000 years old). Population is about 65,000 inhabitants (Medium-sized city). Developed around the 19th- century railway system; over 100 years of political and industrial history.

Due to State of Minas Gerais specifics the city chose frugal innovation, doing more with less in its moving to agenting government. There are three frugal innovation cases: 1) carta de servicios – validation and screening of public services; 2) joseias2.0 – AI Agent for Server Support in the implementation of the electronic information system; 3) chochat+ was created for personalized and immediate support for IT questions and ticket management. Those are agentic artificial intelligence cases applied within a municipal government context. The initiative involves the development of AI agents designed to provide specialized support to public servants across a wide range of operational demands, including ICT assistance, guidance on administrative procedures, and enhanced use of government information systems. By automating routine support activities, these agents enable public employees to dedicate more time to higher-value and more complex tasks.

19 JUNE 2026 SESSION 1

Both User and Citizen: Data Protection and Data Collaboration in the Co-Production of Public Services

Zhangyu Ge, doctoral student, School of Public Administration, Huazhong University of Science and Technology, Wuhan, China

Personal data sharing is not merely about overcoming legal and technical barriers to access services, but also involves actively engaging in data collaboration to create public value. Research Question is as follows: How to move beyond the traditional paradigm that views privacy trade-offs as a personal data sharing behavioral mechanism, and instead construct a more holistic theoretical framework to analyze data value co-creation in public services?

The study employs an explanatory mixed-methods design. Quantitative methods will explore the factors influencing data sharing using PLS-SEM. This will analyze the relationships among variables to draw initial conclusions. A total of 1,165 valid questionnaires were collected for the data analysis. The second phase includes semi-structured interviews are employed to collect data from both patients and administrators. This help comprehensively discusses key

findings, enriching the practical value. The field researches were conducted in Luoyang City, Henan Province, and Wuhan City, Hubei Province.

The research provide a citizen-centered theoretical perspective for explaining individual behavior in digital public service contexts more inclusively. The complex nature of personal data-sharing behavior in the public service domain, which involves both sensitivities and externalities, has been systematically observed to some extent. The empirical findings suggest that the focus of digital public service reform should shift from privacy regulations compliance and data extract to fostering a collaborative and mutually beneficial data ecosystem. This article offers government institutions and managers insights into balancing the claims of privacy protection and public participation intention to realize public service co-production.

Digital Platforms in Public Administration: Enabling Intelligent Governance through Data Integration and Institutional Transformation

Abdullaeva Fazilatkhon, bachelor Student, MGIMO University, Russia

The ongoing digital transformation of public administration is increasingly characterized by a shift from fragmented e-government systems toward integrated, platform-based models. While early stages of digital government primarily focused on digitizing public services, contemporary developments emphasize the creation of digital platforms as foundational infrastructures that enable more advanced and intelligent forms of governance. However, despite the growing importance of digital platforms in public administration, their role as institutional infrastructures enabling intelligent governance remains insufficiently conceptualized in existing research. This research critically examines the role of digital platforms in reshaping public administration and analyzes how platform-based architectures contribute to the emergence of intelligent governance.

Intelligent governance is understood as the capacity of public institutions to utilize integrated data, advanced analytical capabilities, and adaptive decision-making mechanisms. In particular, the study focuses on three core mechanisms: data integration across governmental agencies, interoperability between information systems, and the development of advanced analytical capabilities. Together, these elements allow governments to move beyond reactive service delivery toward proactive, data-driven decision-making processes that enhance efficiency, responsiveness, and policy effectiveness. At the same time, the technological innovation alone is insufficient to achieve intelligent governance. The successful implementation of digital platforms depends on broader institutional transformation, including the establishment of coherent data governance frameworks, the development of regulatory and legal mechanisms, and the enhancement of organizational and human capacities within public institutions. Without these complementary changes, digital platforms may fail to deliver their expected benefits and

can even generate new risks, such as increased centralization, data asymmetries, and institutional fragmentation. The study makes two key contributions. First, it conceptualizes digital platforms as institutional infrastructures of intelligent governance. Second, it identifies the institutional and technological conditions under which platformization leads to effective transformation in public administration.

Integral Assessment of the Effectiveness of Industrial Goods Promotion as a Tool for Intelligent Management in the Digital Economy

Maxim Kravchenko, doctoral student, Russian Presidential Academy of National Economy and Public Administration, Russia

The digital transformation of industry affects not only production processes but also approaches to managing product promotion in the market. In the digital economy, the effectiveness of industrial goods promotion is determined not by isolated marketing indicators but by the coordinated functioning of interconnected elements, including digital communications, sales, logistics, service support, and data management. The research problem lies in the fact that many industrial enterprises evaluate promotion performance using separate indicators such as leads, conversion rates, sales volume, or marketing expenditures. However, this approach does not fully reflect the specifics of industrial business-to-business markets, where overall performance also depends on delivery reliability, order fulfillment speed, service quality, and the level of business process digitalization.

The purpose of the study is to develop an integral approach to assessing the effectiveness of industrial goods promotion in regional markets. The proposed model integrates four analytical components: communication, commercial, logistics-service, and digital. This approach makes it possible to obtain a unified assessment of promotion effectiveness while considering the influence of multiple factors on overall performance. The methodological framework of the study is based on systems, process, and index approaches. The integral indicator is formed through the normalization of variables with different directions of influence, ensuring the comparability of heterogeneous indicators and their integration into a single assessment framework.

The practical significance of the research lies in the possibility of using the integral index as a tool for intelligent management. The proposed approach enables managers to identify weaknesses in the promotion system, evaluate the contribution of individual factors, monitor performance dynamics, and justify managerial decisions related to the development of digital channels, logistics, service operations, and data analytics. The study concludes that the integral assessment of promotion effectiveness contributes to the development of a comprehensive data-

driven management system and supports the competitiveness of industrial enterprises in the digital economy.

The Image of Authority in Digital Format: How Citizens Perceive the Names of Chatbots for Public Services?

Agnessa Bich, master student, HSE University, Saint Petersburg, Russia

Trust in government chatbots is essential for citizens' willingness to use digital public services, yet the role of chatbot names as an initial trust cue remains understudied. This article examines naming as a signal of trust and legitimacy in digital government. We argue that, in public-sector contexts, humanized or playful names do not necessarily enhance trust because citizens approach state institutions with different expectations. Using a sequential mixed-method design, we developed a replicable onomastic typology from a corpus of 315 chatbot names used by Russian public authorities and tested the effects of these naming strategies in a perception survey. Research questions were examined in a Russian sample (N = 151) and subsequently assessed in a near-cultural replication across nine post-Soviet countries (N = 105). Findings indicate that functional and institutional names generate the highest expectations of trust, legality, cybersecurity, and usefulness, whereas anglicized tech-styled and playful names evoke greater skepticism. Anthropomorphic names increase perceived warmth but not institutional trust. The study contributes to socio-onomastic research and digital government scholarship by demonstrating that chatbot names operate as pragmatic signals of legitimacy. In public-sector settings, trust is built more effectively through clarity, role transparency, and institutional alignment than through interpersonal simulation.

Artificial intelligence in public administration: balancing efficiency and ethical norms

Anastasya Sviridova, student, Siberian Institute of Administration, Russian Presidential Academy of National Economy and Public Administration, Russia

The relevance of the topic is due to the active integration of artificial intelligence technologies into the work of government bodies worldwide. AI is seen as a tool for increasing the efficiency of administrative processes, optimizing resource allocation and improving the quality of public services for citizens. The potential of AI in public administration includes automation of routine operations, predictive analytics for decision-making, personalization of services and enhancement of control functions. Examples of successful implementation demonstrate reductions in costs and processing times.

The main risks are associated with ethical aspects: algorithmic bias, discrimination of certain groups of citizens, lack of transparency of neural networks ("black box"), as well as the erosion of accountability in case of erroneous decisions made on the basis of AI

recommendations. The key challenge lies in finding a balance between technological efficiency and compliance with ethical norms. This requires the development of clear legal frameworks for the use of AI, the implementation of algorithmic explainability standards and the mandatory retention of human control over socially significant decisions. The report analyzes international regulatory experience of AI in the public sector and proposes criteria for ethical implementation such as personal data protection, inadmissibility of full automation in sensitive areas (courts, social benefits) and public audit of algorithms.

19 JUNE 2026 SESSION 2

The Digital State and the Fragmentation of the Global Digital Economy: Institutional Models and Digital Sovereignty

*Ekaterina Gurova, Assistant chair (practitioner teacher) World Economy Department,
Saint Petersburg State University (SPbSU)*

The purpose of this report is to analyze digital government models in the context of the transformation of the global digital economy, as well as to identify their impact on the fragmentation of digital space and the formation of technological inequality. The comparative analysis allows us to identify three basic models of digital government:

1. The liberal model of USA is characterized by the dominance of market mechanisms and private digital platforms. Within this model, the state performs primarily regulatory functions, while key digital infrastructures are developed by the private sector. This ensures high innovation dynamics and global competitiveness of platforms, but is accompanied by a concentration of market power and limited state control over data and digital infrastructure. This model operates primarily according to bottom-up logic, where digital transformation is initiated from below – by technology companies.

2. The regulatory model of European Union is based on the principles of socially oriented market regulation. The state acts as an institutional arbiter, shaping the regulatory environment for the functioning of digital markets. Regulatory instruments (GDPR, DMA, DSA, AI Act) play a central role, ensuring data protection, competition, and limiting market concentration. This model combines bottom-up innovation with strong top-down regulation, creating a system of «sovereignty through rules» and exerting global regulatory influence («Brussels Effect»).

3. The centralized model in China is based on the principles of state capitalism and is characterized by the dominance of top-down logic. The state acts as the key coordinator of digital transformation, controlling infrastructure, data, and strategic platforms. This ensures high manageability and speed of implementation of digital projects, but comes with limited autonomy for the private sector and reduced market competition.

The hybrid model of Russia combines elements of centralized governance and market mechanisms. The state determines strategic directions for digital development and creates the infrastructure base, while the private sector maintains a significant role in developing digital solutions and services. This model is characterized by a transition from infrastructural digitalization to institutional transformation of public administration and a stronger focus on digital sovereignty. The analysis shows that the differences between digital government models are largely determined by the balance between top-down and bottom-up mechanisms. A dominant bottom-up approach ensures innovative leadership but reduces the degree of state control over strategic resources. Top-down models, on the other hand, enhance governance and strategic coordination but can limit innovative flexibility.

A key outcome of digital transformation is the emergence of digital sovereignty, defined as the state's ability to control data, platforms, and infrastructure. This is leading to a transition from a globally integrated digital economy to a system of «digital jurisdictions» within which different regulatory and technological regimes operate. A comparison of liberal, regulatory, centralized, and hybrid models of digital government reveals that their differences stem from a combination of top-down and bottom-up institutional logics. The dominance of the bottom-up approach ensures high innovation dynamics and the development of global platforms, while top-down models strengthen state control, strategic coordination, and the ability to implement large-scale digital transformations. Hybrid models strive for a balance of these approaches, but their sustainability depends on institutional maturity and external development conditions.

The strengthening of digital sovereignty is transforming the global digital economy from a single, integrated system into a collection of partially interoperable digital jurisdictions. This is accompanied by fragmentation of the digital space at the regulatory, technological, and economic levels, as well as increased competition between digital government models. At the same time, technological inequality is widening, driven by differences in the digital maturity of states. This results in persistent asymmetries in access to digital resources, technologies, and platforms, which impact countries' positions in the global digital hierarchy. Thus, the further development of the digital economy will be determined not only by technological innovations, but also by the institutional capacity of states to form effective digital governance models that ensure a balance between innovative development, regulation, and digital sovereignty.

The Problem of Trust in Legal Artificial Intelligence Solutions in the Context of Ensuring Human and Civil Rights and Freedoms

Makovskaya Polina, Lawyer, Rostov Regional Collegium of Advocates "Business and Law", Rostov-on-Don, Russia

Specialized artificial intelligence (AI) systems need different principles of use for their effective application, depending on the area of their "vital activity". Rational principles of using AI are of particular importance as a way to ensure human and civil rights and freedoms during the digitalization of the state legal sphere. By the term "legal artificial intelligence" we mean a set of technological solutions that allow us to simulate human cognitive functions (including searching for solutions without a predefined algorithm) and obtain results when performing specific tasks designed to support, support, or perform certain actions in the legal field, functioning with a well-defined degree of autonomy, ensuring transparency of its conclusions, compliance with the principles of justice and legal protection, as well as subject to special certification and control in order to minimize risks to the rights and legitimate interests of participants in legal relations¹.

The use of legal artificial intelligence clearly exposes the problem of trust in the solutions offered by the machine, in the light of which we can talk about the fundamental problems of the relationship between man and the mechanism, in which man appears as *homo credulus*: a person inclined to trust (not to be confused with *homo credens* – a person of faith). The main feature of *homo credulus* is the forced need to trust those devices and objects that ensure its existence and even life, which directly concerns ensuring human and civil rights and freedoms in a "digital" state.

Significant progress in assessing the reliability of AI systems can be achieved by an approach that converts a single reliability metric into a system of four parameters. 1) Consistency, repeatability – the same task under the same conditions should give a comparable result; 2) robustness, resistance to deviations from ideal conditions associated with information noise, poorly recognized photocopy of a document, etc.; 3) predictability, calibration of confidence and prevention of erroneous self-confidence - the ability of the system to say "I don't know" or "I'm not sure"); 4) safety, limitation of harm caused by an error². It should be noted that high values of the predictability parameter can paradoxically cause *homo credulus* to increase confidence in AI solutions. This is confirmed by relevant research: if an AI system honestly "admits" to its lack of confidence in solving certain tasks, this creates a higher integral trust in such artificial intelligence.

¹ Ovchinnikov, A. I., Makovskaya, P. N., Fatkhi, V. I. (2025) Principles of Ethical and Legal Regulation of Artificial Intelligence Technologies in Education and Scientific Research. Bulletin of the Law Faculty, SFEDU. 12(4): 52–62 [in Russ.]. DOI: 10.18522/2313-6138-2025-12-4-6

² Rabanser S. et al. Towards a Science of AI Agent Reliability. arXiv: 2602.16666v2, 2026, 23 February.

Tools for Protecting Against Reward Function Hacking and Systemic Corruption in Artificial Intelligence of Government Systems

Dojdikov Anton, Senior Research Fellow Institute of Socio-Political Research of FCTAS

RAS, Moscow, Russia

In the context of the digital transformation of public administration, reinforcement learning-based artificial intelligence is increasingly embedded in governmental information systems, resource allocation mechanisms, and decision-making circuits. At the same time, multi-agent deep learning models that optimize formal proxy metrics often maximize reported indicators rather than socially meaningful outcomes, thereby generating risks of reward hacking and new forms of systemic corruption in the digital environment. Such distortions may manifest themselves as a persistent divergence between the growth of formal performance indicators and the actual dynamics of social welfare, institutional quality, and the efficiency of resource use.

Methodologically, the study rests on the concept of computational law, which allows for the formalization of state-legal and political processes and their modelling in an algorithmic environment, as well as on the Constitutional AI approach, under which normative principles and constraints are embedded directly into the agent's learning architecture. I propose a multi-agent model of the "federal center-regions" type, in which agents take decisions on budget allocation and implementation of infrastructure projects while facing a conflict between maximizing formal indicators and achieving long-term strategic goals. Constraints in the spirit of Constitutional AI are implemented as elements of the reward function that emulate legal and oversight mechanisms: sanctions, audit procedures, thresholds of socially acceptable outcomes, and the priority of public welfare over local gains of individual agents. The aim of the study is to empirically test whether such embedded constraints can reduce the intensity of reward hacking and mitigate manifestations of systemic corruption in a multi-agent environment. Two regimes of deep reinforcement learning based on the MAPPO algorithm are compared: a baseline regime without constraints and a regime with explicitly specified constraints incorporated into the reward function. Each regime is modelled in a series of 20 independent runs of 200 episodes. Non-parametric statistical tests are used to assess differences between the regimes, which makes it possible to establish the robustness of the observed effects.

The results show that the introduction of constraints leads to statistically significant changes in system behaviour. In the second regime, global social welfare increases by 46.19 conditional units, the reward hacking index decreases by 0.097, the gap between proxy metrics and target outcomes shrinks by 20.97, the institutional distortion index decreases by 0.257, and the aggregate level of violations decreases by 2.20. These findings confirm the hypothesis that architectural constraints embedded in the reward function are capable of shifting agent

optimization away from manipulating formal indicators towards achieving the genuine goals of public policy. The scientific contribution lies in the conceptual convergence of the theory of systemic corruption and the phenomenon of reward hacking in artificial intelligence systems, as well as in the interpretation of the Constitutional AI approach as an imperative architectural model in which legal principles and accountability mechanisms take the form of computational constraints. The practical significance of the study is that the proposed model allows internal architectural parameters of AI systems to be viewed as an additional layer of legal and institutional control, which is essential for the development of new regulatory requirements for governmental and commercial information systems based on deep multi-agent reinforcement learning.

When an Algorithm Evaluates People: AI-Assisted Management and the Limits of Algorithmic Assessment

Gracheva Alina, Project Manager, «The Best App» LLC, Nizhny Novgorod, Russia

The research question is what happens when an algorithm not just monitor labour, but starts to assess it? Management instruments are evolving from passive data registration to the active evaluation. The end-to-end case is the platform for project management “Logwork”. It is a company’s creation, which analyses the digital footprints of employees. Digitalization is not just about software implementation. This is a qualitative change in the distribution of cognitive functions. "Reassigning to computer algorithms partially or completely the functions previously assigned to human decisions"³. Digitalization begins where an algorithm takes over part of the cognitive load, rather than just storing data.

The first stage is traditional task management: manual task assignment, status updates through communication, and reporting that fully depends on employees’ self-reports. Tools such as spreadsheets or early task trackers served as containers for information, but they did not create any analytical layer. The second stage is routine automation: systems based on predefined rules that reduce the burden of repetitive communication. Examples include Telegram bots with deadline reminders, automatic dashboards that aggregate working time, and team workload heatmaps. The key difference is that automation follows pre-programmed logic: it does not learn and does not make conclusions. The third stage is AI analytics: systems that not only collect data, but also generate its interpretation. This evolution follows the logic described by Mikhnenko in the context of project change management. According to the author, the modern project environment requires the “analysis of multimodal data”, that is, the simultaneous processing of different types of sources, including texts, numerical indicators, and activity

³ Shabaltina L. V., Maslennikov V. V. Management of digital transformation of organizations using artificial intelligence. Issues of innovative economics. — 2023. — Vol. 13, No. 2. — pp. 771-784. DOI: 10.18334/vinec.13.2.118231

records, to support managerial decisions. Mikhnenko also notes that the most in-demand AI technologies in project management are natural language processing and intelligent decision support. These technologies make it possible to move from data aggregation to data interpretation⁴.

Activity Summary module “Logwork” can make artifact analysis: commits to Gitlab and layout changes to Figma – the system scans the digital footprint automatically. Text summary – “Logwork” generates a description of each employee's activity based on the collected data. It also highlights anomalies, automatically highlights "what needs attention": absence of an artifact, deviations in time deduction. This represents the shift from data aggregation to evaluation.

Three reasons why algorithmic assessment may not be objective:

1. Blindness to invisible work: AI sees only a digital footprint. Commit does not show the quality of work. Analysis, documentation, and conflict resolution are invisible to the algorithm.

2. Goodhart's Law: A measure that has become a goal ceases to be a good measure (Goodhart, 1975; Strathern, 1997). Employees optimize the metric, not the result.

3. Erosion of managerial judgment. A manager who relies only on AI reports loses the skill of independent interpretation (Vigoda-Gadot, 2024).

AI reports are perceived as neutral, however, this is an illusion of objectivity. Algorithmic evaluation always reflect the assumptions embedded in it. The main question should be for what decisions and to what extent should we trust the algorithm? Relevance for public administration is that risks are the same – opacity, "black box", inability to challenge. The key conclusion is that awareness of use is the foundation of intelligent management in the digital age

⁴ Mikhnenko P. A. Analysis of multimodal data in project management: prospects for using machine learning. Management Sciences. — 2023. — Vol. 13, No. 4. — pp. 71-89. DOI: 10.26794/2404-022X-2023-13-4-71-89.