RESEARCH PAPERS ON KNOWLEDGE, INNOVATION AND ENTERPRISE
VOLUME 14, ISSUE 1, JULY 2026, PP. 135-151

AI Protective Models as a National Content Security Infrastructure: Russian Experience

ALEXEY A. KUZMIN
AI SenseSafety Center, IRIAS, Moscow, Russia

Abstract

Large Language Models (LLMs) are increasingly embedded in public, commercial and regulated services, making operational content safety a matter of infrastructure rather than model design alone. This paper argues that value-alignment techniques, while important, cannot by themselves ensure that deployed LLMs comply with the legal and cultural requirements of specific jurisdictions. It therefore introduces the concept of an AI protective model: a specialised, independently trained and certified safety component deployed as an API/SDK interceptor within the LLM inference pipeline. The proposed model operates across two complementary dimensions: legal compliance (LC), which adapts content controls to jurisdiction-specific legal classifications, and Cultural Code Protection (CCP), which addresses culturally harmful content that may fall outside explicit legal prohibitions. The paper develops the theoretical basis for CCP, proposes operational criteria for distinguishing cultural sovereignty from censorship, and outlines a three-tier institutional architecture for the AI SenseSafety Center (AISSC), comprising protective-model development, independent certification and deployment through cloud or isolated on-premise environments. Particular attention is given to the emerging Russian regulatory context and to the potential use of the Guard Crosswalk concept for future inter-jurisdictional certification alignment, especially across SCO, BRICS and ASEAN contexts. The paper concludes that AI protective models should be treated as a distinct layer of national digital infrastructure, complementary to value alignment and capable of independent updating, auditing and certification. It also identifies the need for empirical validation, transparent governance and further international dialogue.

Keywords: AI protective models, AI safety, legal compliance, Cultural Code Protection (CCP), AI governance, national digital infrastructure

Introduction

Large Language Models (LLMs) have evolved from experimental systems into production infrastructure. Government portals respond to citizens' requests, clinical systems support diagnostics, and educational platforms serve millions of students – all using LLMs in continuous inference mode. According to McKinsey Global Institute (2025), by 2025 more than 60% of Fortune 500 companies had implemented generative AI in at least one production process. The rapid expansion of such deployments means that AI content safety is no longer primarily an academic concern, but an engineering and legal one.

The global community has responded to this challenge primarily through methods of model value alignment. RLHF (Reinforcement Learning from Human Feedback) (Stiennon et al., 2020), Constitutional AI (Bai et al., 2022), Direct Preference Optimization (Rafailov et al., 2023) and related approaches together form what may be described as the "built-in values" of a model: they shape how the model tends to respond under normal conditions. The significance of this work is reflected in measurable results: state-of-the-art LLMs are substantially less likely to generate obviously harmful content than their 2021–2022 predecessors.

However, value alignment addresses a fundamentally different task from the one considered in this article. Value alignment asks: is the model intended to cause harm? Operational AI content safety asks a different question: does the model's behaviour comply with the specific legal norms of a particular jurisdiction, in live deployment, at this moment? These are distinct tasks. Even a perfectly aligned model does not necessarily know what is classified as prohibited content under Russian, Chinese, Indian or German law. Empirically, even the best-aligned models retain exploitable vulnerabilities, as numerous independent studies using adversarial analysis have shown (Kizilcec et al., 2024; Mazeika et al., 2024).

A second, less formalised dimension sits alongside this task: beyond clearly illegal content lies a broad "grey zone" – content that is not explicitly prohibited but may nevertheless cause harm from the standpoint of a community's values. This zone is not adequately covered either by value-alignment techniques or by existing commercial solutions, which are oriented primarily towards the legal environment of the United States.

This article proposes a conceptual and institutional solution to both tasks. It introduces the concept of an AI protective model – a separate architectural component acting as an API/SDK interceptor within the LLM inference pipeline – and describes the three-tier institutional model of the AI SenseSafety Center (AISSC), established at the International Research Institute for Advanced Systems (IRIAS, Moscow). The article is structured as follows. Sections 2–3 analyse related solutions and introduce the concept of an AI protective model; Sections 4–5 develop its theoretical and institu-

tional foundations; and Sections 6–7 address the regulatory dimension and the limitations of the approach.

Literature Review and Related Solutions

Value alignment: achievements and limitations

The RLHF method proposed by Stiennon et al. (2020), and subsequently systematised in the work of OpenAI and DeepMind, has become the dominant approach to shaping LLM behaviour in line with human preferences. Constitutional AI (Bai et al., 2022) made it possible to move from labelling individual responses towards explicitly specifying the principles that a model should follow. Later methods – including Direct Preference Optimization (Rafailov et al., 2023) and RLAIF – reduced dependence on large-scale human labelling while preserving the effectiveness of preference-based alignment.

The achievements of value-alignment techniques are substantial: state-of-the-art LLMs in 2025 produce a markedly lower proportion of undesirable responses on standardised evaluation tests than models from the 2021–2022 generation. At the same time, a growing body of empirical evidence points to three structural limitations of value alignment when it is used as the sole AI-safety mechanism.

First, residual vulnerabilities. Even well-aligned models remain susceptible to jailbreak-type attacks. Kizilcec et al. (2024) documented persistent cultural biases in GPT-4 and comparable systems: models systematically orient themselves towards US value systems, reflecting the underrepresentation of non-Western jurisdictions in RLHF training data. Attacks such as AutoDAN (Liu et al., 2023), Tree of Attacks (Mehrotra et al., 2023) and GCG achieve success rates of 60–94% against aligned models when applied automatically (Mazeika et al., 2024). Independent studies have also confirmed that harmful value patterns can be reproduced even under explicit value alignment (Kizilcec et al., 2024; Mazeika et al., 2024).

Second, jurisdictional blindness. Value-alignment training takes place within the legal environment of the developer – typically the United States or the EU. Yet classifications of prohibited content under Russian, Chinese, Indian and Brazilian law do not necessarily correspond to those legal coordinates. An aligned model "does not know" what constitutes an offence in a particular jurisdiction not because it has "wrong values", but because detailed legal classifications cannot reliably be embedded in pre-training parameters: legislation changes over time.

Third, dependence on an external developer. Even where an operator wishes to strengthen safety, it cannot modify the weights of a foreign-developed base model. Updating safety-related behaviour therefore requires a full cycle of retraining or fine-tuning – a procedure that is unavailable to most operators of AI services.

Existing commercial "guard" solutions

Leading vendors offer solutions that partially address content filtering at the level of LLM inputs and outputs. Meta Llama Guard (Inan et al., 2023) is a Llama-based classifier trained on the MLCommons taxonomy of harmful categories. Amazon Bedrock Guardrails implements input/output filtering through configurable policies. NVIDIA NeMo Guardrails provides programmable dialogue flows governed by user-defined rules. Microsoft Azure AI Content Safety offers an API service based on a proprietary classification model.

These solutions share several critical limitations: proprietary and non-transparent architectures and training data; jurisdictional orientation towards the United States; no option for fully isolated deployment for government and regulated clients; no independent certification against national or international AI-governance standards; and technological dependence, since their use may require data to be sent to foreign cloud operators. As a result, none of the existing commercial solutions fully addresses the needs of regulated environments in jurisdictions outside the United States.

AI safety institutions: mandate and boundaries

Since 2023, a network of national AI Safety institutions has emerged, including the UK AISI (AI Security Institute), US AISI, J-AISI (Japan), CAISI (Canada), the EU AI Office and several others. Their collective contribution is substantial: systematic testing of frontier LLMs under attack, development of TEVV (Testing, Evaluation, Validation, Verification) methodologies, research on interpretability and elicitation, and the shaping of international standards through ISO/IEC JTC 1/SC 42.

Their mandate, however, is centred on frontier LLMs during development and pre-release evaluation. Operational AI content safety in production, the jurisdictional compliance of specific deployments, and the certification of safety components for particular legal systems largely fall outside this mandate. As stated in the UK AISI 2025 Annual Report (UK AI Security Institute, 2025): "our task is to assess systemic risks on the horizon, not to ensure that each particular deployment is compliant with national law".

The distinction is fundamental: even where an international AI-safety infrastructure is mature, the jurisdiction-specific operational layer can remain unaddressed. It is precisely this gap that AISSC is designed to fill.

Regulatory context

International AI-governance frameworks establish requirements for operational safety but generally do not prescribe specific technical implementations. NIST AI RMF (NIST, 2023) defines the functions Govern, Map, Measure and Manage while leaving the choice of tools to the organisation. The EU AI Act (Regulation (EU) 2024/1689) introduces mandatory risk-

management (Art. 9) and accuracy, robustness and cybersecurity (Art. 15) requirements for high-risk AI systems without specifying the architecture of protective components. ISO/IEC 42001 (ISO/IEC 42001:2023, 2024) sets requirements for AI management systems at the organisational level.

A fundamentally new development is the draft Federal Law of the Russian Federation "On the Fundamentals of State Regulation in the Sphere of Artificial Intelligence Technologies", published on regulation.gov.ru and currently being revised by the Ministry of Digital Development, Communications and Mass Media of the Russian Federation (Ministry of Digital Development, 2026). Articles 8–10 of the draft introduce AI protective models as a mandatory component of the AI-service lifecycle and envisage an authorised national operator for their certification. To the author's knowledge, this is the first instance in which protective models have been enshrined as a distinct category in a draft national AI law.

The Concept of an AI Protective Model

Definition and architectural principle

An AI protective model is a specialised AI-safety component deployed as an API/SDK interceptor within the LLM inference pipeline and architecturally separated from the base model; it is trained, certified and versioned independently.

The architectural logic of this definition draws on a principle well established in network security: the firewall principle. A firewall is not built into the protected application; it functions as a separate security layer that can be updated and audited independently. This separation is precisely what makes the architecture robust: security does not depend on the internal structure of the protected system.

AI protective models implement the same principle. The base LLM may be of any origin: domestic (GigaChat, YandexGPT) or foreign (GPT-4, Claude, Llama), open or proprietary. The protective model does not modify the base LLM; instead, it operates as a separate layer, intercepting requests (input control) and/or responses (output control) at the inference-infrastructure level. It follows that:

- the service operator remains independent of the base-model developer's safety policy;
- the protective model can be updated when legislation changes without retraining the base model;
- the protective model can be independently certified as a standalone component;
- where necessary, the protective model can be deployed in a fully isolated environment.

Two dimensions of protection

An AI protective model implements two independent but complementary dimensions of protection.

The first dimension is legal compliance (LC). The protective model ensures that LLM behaviour conforms to the norms of a specific jurisdiction: classifications of prohibited content, restrictions on information distribution, age-rating requirements, personal-data provisions and other legal prescriptions. The key property of this dimension is jurisdictional specificity: a protective model for Russian deployment is trained on Russian legal classifications; for EU deployment, on relevant EU taxonomies, including those associated with the AI Act and DSA. The same base LLM can therefore serve several jurisdictions simultaneously through different protective models.

The second dimension is Cultural Code Protection (CCP). A protective model prevents users from receiving content that is harmful from the standpoint of their community's values, regardless of whether corresponding legal regulation has been established. This concerns the "grey zone": content that is not directly prohibited by law but may inflict value-level harm on a particular cultural community. The theoretical foundations of this dimension are developed in Section 4.

The two dimensions are independent in the sense that each can be activated separately. The operator can choose among several configurations: a universal protective model (both dimensions simultaneously), a specialised legal-compliance model (LC only), or a specialised CCP model (CCP only). This flexibility allows the level and type of protection to be adapted to the operator's specific regulatory and value context.

Delivery model and the Protective Models Library

AI protective models are delivered in two deployment modes: a cloud service (SaaS) via API and on-premise (isolated) deployment. The cloud mode is intended for commercial AI-service operators that do not handle state secrets or other restricted data. On-premise deployment is intended for public authorities, critical information infrastructure and operators working with personal data in secured environments.

Protective models are managed through the Protective Models Library – a versioned registry of certified AI protective models with open access to metadata, including domain of application, version, certification date and applicable standards. Each protective model in the Library is accompanied by a certification card describing its training data, test methodology and results of attack-resilience evaluation.

Cultural Code Protection: Theoretical Foundations

Cultural code as a scientific category

The notion of cultural code has deep roots in anthropology, cultural studies and semiotics. Claude Lévi-Strauss described cultural codes as systems of

distinctions through which a community organises its experience. Clifford Geertz defined culture as "webs of significance" in which humans live (Geertz, 1973). In a more formalised form, Geert Hofstede operationalised cultural differences through measurable dimensions – including power distance, individualism/collectivism and uncertainty avoidance – creating a widely used tool for comparative analysis, including studies of LLMs. It was precisely through Hofstede's framework that Kizilcec et al. (2024) documented cultural bias in state-of-the-art LLMs.

It is important to stress that Hofstede's model is used here solely as a diagnostic tool – a framework that enables independent researchers to quantify an already observed phenomenon of cultural bias in LLMs. The two- to six-dimensional Hofstede scheme is not the theoretical basis for CCP: the cultural code addressed by AI protective models is inherently high-dimensional and cannot be reduced to a small number of fixed axes.

For the purposes of this article, cultural code is understood as the set of stable value orientations, narrative patterns, communication norms and symbolic systems shared by a community and reproduced across generations through language, education, institutions and practices. This definition is operational rather than philosophical: it identifies what can be formalised, labelled and used to train AI protective models. At the same time, the concept is used broadly to cover both stable value configurations within a community (its current cultural profile) and the rules governing their reproduction and transformation over time (cultural dynamics). A rigorous separation of these two aspects – the static cross-section and the dynamic operator of cultural code – is a subject for separate theoretical work and lies beyond the scope of this engineering-oriented article.

Cultural code must also be distinguished from political preferences. Cultural code is a stable, long-term characteristic of a community that can be described through scientific methods, including comparative cultural studies, sociology of values and corpus linguistics (Johansen et al., 2025). Political preferences are situational, shifting and contested. CCP rests on the former and explicitly excludes the latter: a protective model should follow scientifically validated descriptions of cultural code, not current political agendas. Cultural code as an engineering category intentionally brings together several levels of cultural reality – symbolic, value-based, normative and communicative – without claiming an ontological distinction between them. Determining where these levels lie, which are amenable to formalisation and which are not remains a separate theoretical challenge for the school of formal cultural modelling.

Formalisation for machine-learning tasks

Translating the notion of cultural code into an operational format suitable for training AI protective models requires work at three levels.

The first level is taxonomy. A hierarchical classification of content categories must be constructed from the standpoint of cultural code: neutral, potentially sensitive, and clearly violating community cultural norms. This taxonomy is developed through expert consensus involving cultural scholars, sociologists, scholars of religion and representatives of the educational community – rather than political bodies.

The second level is a labelled training set. The taxonomy is operationalised as a labelled training dataset containing examples of content with a justified classification for each category. The training set is subject to open audit regarding its formation methodology, labelling criteria and procedures for resolving disagreements among annotators.

The third level is verification and drift. Cultural code is not static: community values evolve. The CCP protective model must therefore be revised periodically on the basis of monitored sociological indicators, for example World Values Survey Association data (2022). Versioning in the Protective Models Library ensures transparency of change: each new model version is accompanied by a report describing changes to the training set and justifying the adjustments.

Boundary between cultural sovereignty and censorship

The most controversial aspect of CCP requires a clear operational criterion. The author proposes the following distinction.

Cultural sovereignty is realised when: (a) a restriction is based on a verified description of community values obtained through scientific methods; (b) the restriction applies to content rather than to the subject of the request; (c) the procedure for forming restrictions is transparent and open to independent audit; and (d) restrictions are regularly revised on the basis of sociological data.

Censorship is realised when: (a) a restriction is driven by political exigency rather than verified cultural norms; (b) the restriction targets specific subjects rather than content categories; and (c) the procedure for forming restrictions is opaque and not subject to audit.

This criterion functions as an engineering constraint for CCP protective-model development: it excludes from the training set categories that do not meet conditions (a)–(d). CCP is intended to preserve cultural identity rather than enforce political control, and the engineering criterion is what makes this distinction operational rather than merely declarative.

Relationship to international legal frameworks

The CCP concept operationalises international human-rights standards in the AI environment. The UNESCO Universal Declaration on Cultural Diversity (UNESCO, 2001) explicitly recognises communities' right to protect cultural expression. The UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions (UNESCO, 2005) establishes the right of states to adopt measures to protect cultural diversity,

including in the digital environment. The Johannesburg Principles on National Security, Freedom of Expression and Access to Information (1996) distinguish between legitimate limitations on freedom of expression and abuses of such limitations.

Applied to CCP, this implies that states not only have the right but also an obligation to maintain conditions for preserving communities' cultural codes in the digital environment – a function that this dimension of AI protective models is designed to support.

Institutional Architecture of AISSC

Rationale: why a national operator is needed

Providing operational AI content safety at jurisdictional scale requires an institutional actor that combines several characteristics rarely found together: deep AI technical expertise, understanding of national legal regulation, capacity to work with data in secured environments, and regulatory authority for certification.

Neither commercial AI developers (technically competent but jurisdictionally limited and procedurally opaque), regulators (legally competent but often lacking comparable technical expertise), nor academic institutions (competent but lacking mandate and delivery infrastructure) fully combine these capabilities. A specialised national operator – an institution of a new type – is therefore required.

The International Research Institute for Advanced Systems (IRIAS, Moscow, Russian Federation) has a distinctive combination of characteristics relevant to this role. Founded in 1976 under the UNECE framework, IRIAS has spent almost fifty years specialising in systems analysis, control theory and applied informatics. The systems approach underlying the IRIAS research school provides a methodological foundation for designing complex AI-safety infrastructure: understanding component interdependencies, managing system lifecycles and standardising complex technical processes. In Russia, an emerging regulatory framework introduces AI protective models as a mandatory component of the AI-service lifecycle and envisages an authorised national operator for their certification (see Section 6). AISSC was established with the intention of fulfilling this function.

Three-tier architecture

AISSC implements a comprehensive approach based on three interconnected tiers.

Tier 1 – Protective models layer. This layer consists of autonomous AI components – AI protective models – developed and maintained by AISSC. Each protective model is trained on jurisdiction-specific datasets (for LC) and culture-specific datasets (for CCP). Development includes mandatory adversarial testing. The architecture is open: protective models are compatible with any base LLM through a standardised API. All models

undergo independent certification before publication in the Protective Models Library.

Tier 2 – Certification layer. At this tier, each protective model undergoes independent evaluation before publication in the Library. The procedure includes verification of the training-set formation methodology, testing for resistance to adversarial attacks, verification of compliance with applicable legal classifications of prohibited content (for LC), and assessment against scientifically validated descriptions of cultural code (for CCP). The core tool at this tier is Guard Crosswalk, designed to map certification results against applicable AI-governance frameworks across jurisdictions. The outcome is a certification card – a public document describing the model's domain of application, version, certification date and applicable standards.

Tier 3 – Delivery layer. This tier is responsible for deploying certified AI protective models in two modes: SaaS (cloud, via API) for commercial operators and on-premise deployment for government and regulated clients. The Delivery layer is supported by the Protective Models Library – a versioned registry providing open access to certification metadata. Planned milestones are Protective Models Library v1.0 (20 base AI protective models) by the end of 2026 and v2.0 (50+ specialised models) by 2028.

Positioning in the AI safety ecosystem

AISSC occupies an institutional niche that is clearly distinct from existing AI-safety actors.

Table 1. *Positioning of AISSC in the AI safety ecosystem*

Characteristic	**AISI (UK / US / J)**	**Commercial vendors**	**AISSC**
Focus	Frontier models, value alignment	Product-level content filtering	Operational safety of LLM deployments
Jurisdiction	Global / country of origin	US / EU	National plus exportable
Certification	No	No	Yes, with mapping to international systems
Isolated deployment	No	No	Yes
Openness	Research-oriented	Proprietary	Open architecture
Cultural dimension	No	No	Yes (CCP)

The central positioning principle is that AISSC does not compete with global AI Safety institutions; it complements them at the operational layer. Value-alignment work shapes a model's "intentions", while AI protective models operated by AISSC are intended to ensure that deployment behaviour is aligned with the jurisdictional and cultural context in which the system operates.

Regulatory Dimension and Inter-Jurisdictional Alignment

AI protective models do not operate in a regulatory vacuum: they are technical tools for meeting operators' legal obligations. Analysis of relevant instruments indicates that protective models can address specific requirements in each framework, although the nature and degree of correspondence vary significantly across jurisdictions.

Russian regulatory priority

The baseline regulatory context for AISSC is Russian AI legislation. A regulatory framework is currently being developed that introduces the concept of AI protective models, defines operators' obligation to apply certified tools to prevent the generation of prohibited content, and envisages an authorised national certification operator (Ministry of Digital Development, 2026). To the author's knowledge, this is the first case globally in which AI protective models are defined in a draft national law as a distinct category subject to mandatory certification. AISSC has designed its institutional architecture with this prospective mandate in mind.

International AI-governance frameworks: analytical context

Alongside Russian law, AISSC engages with international AI-governance systems primarily at analytical and methodological levels.

NIST AI RMF (NIST, 2023) defines the Govern, Map, Measure and Manage functions for AI-system risks. Conceptually, an AI protective model addresses the Manage function for content risks. Risk-assessment methodology from the NIST AI RMF is therefore taken into account in designing the AISSC certification procedure, although any claim of direct compliance will require separate verification.

ISO/IEC 42001 (ISO/IEC 42001:2023, 2024) sets requirements for organisational AI management systems. The architecture of the Protective Models Library – a versioned registry with certification cards – is designed with compatibility with the standard's documentation requirements in mind. The extent to which this compatibility is achieved in practice will be determined during the pilot phase.

The EU AI Act (Regulation (EU) 2024/1689) introduces mandatory risk-management requirements for high-risk AI systems (Art. 9) and requirements for accuracy, robustness and cybersecurity (Art. 15). In designing Guard Crosswalk, these requirements are treated as benchmarks so that

future mapping of Russian certification against the Act's requirements is technically feasible for operators working across Russian and European jurisdictions. At this stage, however, the proposal remains an architectural concept rather than a verified procedure.

Eurasian cooperation as a priority direction

Strategically, AISSC's priority direction for inter-jurisdictional development is the Eurasian space: SCO, BRICS and ASEAN countries that are building their own AI infrastructures under conditions structurally similar to those facing Russia. Many of these states confront comparable challenges: dependence on foreign proprietary content-safety solutions, the jurisdictional limitations of Western guard systems, and insufficient domestic technical capacity to develop alternatives.

The Guard Crosswalk concept is being developed as a potential technical foundation for future Eurasian agreements on mutual recognition of AI-safety certificates. Precedents for mutual recognition in other technically complex domains – pharmaceuticals, aviation and financial regulation – suggest that such mechanisms are feasible where compatible assessment methodologies and mutual trust between institutional operators exist. For Eurasian states, adapting an already developed instrument to their own jurisdiction could involve substantially lower transaction costs than building an analogue from the ground up.

It should be emphasised that Guard Crosswalk is, at this stage, a conceptual development rather than an operational tool. Its practical implementation will require a separate research and engineering cycle, while Eurasian mutual-recognition agreements would require multilateral intergovernmental dialogue. The author therefore presents this article as an argument for the necessity of such dialogue rather than as a description of current practice.

Discussion

Contribution to AI safety architecture

This work proposes a conceptual model that extends AI-safety architecture by distinguishing a standalone operational AI content-safety layer from the value-alignment layer of the base LLM. The distinction is not a demarcation between competing approaches but a functional specialisation within a unified safety system. Value alignment shapes the model's intrinsic tendencies; AI protective models are intended to ensure compliance with the legal and cultural context of a particular deployment without modifying the base LLM.

The analogy with network security operates at several levels. In corporate networks, a firewall is not typically integrated into the endpoint operating system; rather, it is deployed as a separate, independently managed layer. Similarly, the principle of "security as a separate, independently

certifiable layer" is not a compromise but an architectural advantage: it allows independent updating, auditing and replacement without affecting core functionality.

Reproducibility of the model

The institutional model of AISSC is designed as a reproducible template for any jurisdiction that takes AI governance seriously. The minimum necessary components are: (1) a legal classification of prohibited content; (2) a scientifically validated description of cultural code (for CCP); (3) technical capacity to develop and train AI classifiers; and (4) a regulatory mandate for certification.

The first and fourth components exist in many states with mature AI regulation. The second and third require targeted investment but are not uniquely difficult: established methods in comparative cultural studies and the availability of open classifier architectures based on LLMs make the task achievable within 2–3 years for many mid-sized economies.

Limitations and open questions

It is important to state explicitly the limitations of the proposed approach.

Operationalising cultural code. The transition from an anthropological concept to a laelled training set inevitably involves reduction and simplification. Cultural norms are heterogeneous within communities, dynamic over time and often contested. The proposed approach – taxonomy through expert consensus combined with periodic revision based on sociological data – mitigates these risks but does not eliminate them.

Risk of misuse. Any content-filtering tool can potentially be misused. The operational criteria distinguishing sovereignty from censorship (Section 4.3) are necessary but insufficient to prevent abuse. Mechanisms for independent audit, rights to appeal protective-model decisions, and public reporting on use are therefore mandatory elements of the institutional architecture that require further elaboration.

Empirical validation. This work is conceptual: AISSC's architecture is described, but results from production deployment are not yet available. Quantitative evaluation of AI protective models – including precision, recall and false-positive rates – comparative testing against commercial analogues, and adversarial testing of outcomes are tasks for subsequent empirical studies.

International legitimacy. The CCP concept will require broad dialogue with the international AI-safety community and human-rights organisations to validate the proposed criteria of distinction. The author presents this article as an opening to that dialogue rather than its conclusion.

Directions for future research

Based on the limitations identified above, the author proposes the following priority directions for future research:

- development of an evaluation methodology for AI protective models that accounts for jurisdictional specificities;
- creation of open cross-jurisdictional training sets for assessing cultural bias in LLMs;
- research on mechanisms for mutual recognition of AI-safety certificates in the context of SCO, BRICS and ASEAN;
- development of a formal methodology for operationalising cultural code with the participation of cultural scholars, sociologists and AI experts.

Conclusion

This article introduces the concept of an AI protective model and establishes its proposed role as an independent architectural layer within AI safety – complementary to, rather than competing with, value-alignment technologies.

The key contributions are: (1) a formal definition of an AI protective model and justification of the architectural principle separating it from the base LLM; (2) a two-dimensional conceptual model of protection – legal compliance (LC) and Cultural Code Protection (CCP) – with theoretical grounding for both dimensions and operational criteria distinguishing CCP from censorship; (3) a three-tier institutional architecture for the AI SenseSafety Center (AISSC), with the Guard Crosswalk concept as a potential mechanism for inter-jurisdictional certification alignment; and (4) an analysis of the regulatory dimension, with emphasis on the Russian priority and prospects for Eurasian cooperation within SCO, BRICS and ASEAN.

From an AI-policy perspective, the proposed model has one central implication: national AI strategies should include the establishment or recognition of an AI protective-model operator as a mandatory infrastructure component alongside base models, computational infrastructure and training data. Without such a component, an AI-service operator either remains dependent on foreign proprietary filtering or implements ad hoc solutions without a systemic architecture or certification.

AI protective models are not a product or a competitive advantage. They are conceived as a public good within national digital infrastructure, comparable in necessity to antivirus databases, identification systems or legal registries. Precisely in this capacity – as public infrastructure – they require institutional support that cannot be delegated to either academia or the market alone.

Correspondence
Alexey A. Kuzmin
Director, AI SenseSafety Center (AISSC)
IRIAS, Moscow
Russia
Email: a.kuzmin@mniipu.org

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2026). Constitutional AI: Harmlessness from AI Feedback. arXiv, 2022. Available at: https://arxiv.org/abs/2212.08073 (accessed: 10.06.2026).

Geertz, C. (1973). The Interpretation of Cultures. New York: Basic Books.

Hofstede, G. (2001). Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations. 2nd ed. Thousand Oaks: Sage.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., et al. (2026). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv, 2023. Available at: https://arxiv.org/abs/2312.06674 (accessed: 10.06.2026).

Kizilcec, R. F., Tao, Y., Viberg, O., Baker, R. S. (2023). Cultural Bias and Cultural Alignment of Large Language Models. arXiv, 2023. Available at: https://arxiv.org/abs/2311.14096 (accessed: 30.07.2026).

Liu, X., Xu, N., Chen, M., Xiao, C. (2023). AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. arXiv, Available at: https://arxiv.org/abs/2310.04451 (accessed: 10.06.2026).

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv. Available at: https://arxiv.org/abs/2402.04249 (accessed: 10.06.2026).

McKinsey Global Institute (2025). The State of AI in 2025. McKinsey & Company, 2025. Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (accessed: 10.06.2026).

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., Karbasi, A. (2023). Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. arXiv. Available at: https://arxiv.org/abs/2312.02119 (accessed: 30.07.2026).

Ministry of Digital Development, Communications and Mass Media of the Russian Federation (2026). Draft Federal Law 'On the Fundamentals of State Regulation in the Sphere of Artificial Intelligence Technologies'. regulation.gov.ru. Available at: http://regulation.gov.ru (accessed: 30.07.2026).

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023). National Institute of Standards and Technology. Available at: https://www.nist.gov/itl/ai-risk-management-framework (accessed: 30.07. 2026).

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv. Available at: https://arxiv.org/abs/2305.18290 (accessed: 30.07.2026).

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., et al. (2020). Learning to Summarize from Human Feedback. arXiv. Available at: https:// arxiv.org/abs/2009.01325 (accessed: 10.06.2026).

UNESCO, (2025). Convention on the Protection and Promotion of the Diversity of Cultural Expressions. Paris: UNESCO.

UNESCO (2001). Universal Declaration on Cultural Diversity. Paris: UNESCO.

World Values Survey Association (2022). World Values Survey. Available at: https://www.worldvaluessurvey.org/ (accessed: 10.06.2026).