МРНТИ 81.93.29

A.Sinchev[1], B.Sinchev[2], A.Chinchalinova[3*], A.Ospanova[4]

[1]Executive Office of the President of the Republic of Kazakhstan, Astana;
[2]International Information Technology University, Almaty, Kazakhstan;
[3]Executive Office of the President of the Republic of Kazakhstan, Astana;
[4]Digital Government Center, Astana, Kazakhstan.
*E-mail: Aigulch2206@mail.ru

DISTRIBUTED AI ADAPTATION AS A STRUCTURAL RISK VARIABLE: FROM NEURAL ARCHITECTURES TO LAYERED SYSTEM DESIGN

Sinchev Askar – Master's degree, Executive Office of the President of the Republic of Kazakhstan, Astana; Member of the Ethics and Regulation Board of Global Alliance on AI for Industry Centre of Excellence (UNIDO); recognized among the top global innovators in the public sector (Global Innovation Management Institute).

E-mail: askar.sinchev@gmail.com, ORCID: https://orcid.org/0000-0002-7333-2255.

Sinchev Bakhtgerey – Doctor of Technical Sciences, Professor, the Department of Information Systems, International Information Technology University, Almaty, Kazakhstan.

E-mail: sinchev@mail.ru, ORCID: https://orcid.org/0000-0001-8557-8458

Chinchalinova Aigul – Master's degree, Executive Office of the President of the Republic of Kazakhstan, Astana.

E-mail: Aigulch2206@mail.ru, ORCID: https://orcid.org/0009-0001-1482-7784

Ospanova Aliya – Master's degree, EMBA, General Director, Digital Government Center, Astana, Kazakhstan.

E-mail: aliyakalkamanovna@gmail.com, ORCID: https://orcid.org/0009-0006-9977-3696

Abstract. Contemporary artificial intelligence (AI) systems introduce a class of risk that is not event-driven but structurally embedded. Unlike traditional technological risks, which typically manifest through discrete failures or identifiable malfunctions, risks in modern AI systems arise from continuous optimization processes operating over high-dimensional parameter spaces under incomplete constraints. This paper defines such risk as structural adaptation risk, understood as the possibility that system behavior evolves over time in ways that remain internally consistent with the optimization objective, yet diverge from intended functions or policy constraints. This divergence does not require system failure, external interference, or explicit error. Rather, it emerges from the interaction between proxy-based objective functions, distributed representations in multi-layer neural architectures, and adaptive feedback mechanisms, including reinforcement learning and human-in-the-loop evaluation.

Empirical evidence, including cases of specification gaming, reward manipulation, and context-dependent response strategies, indicates that AI systems may systematically optimize for measurable signals rather than intended outcomes. Within this framework, such behaviors should not be interpreted as anomalies, but as consistent outcomes under conditions of incomplete constraint specification. A key implication is that risk may accumulate gradually and become embedded in system behavior before it is externally observable. Consequently, governance approaches based primarily on monitoring, auditing, or post-deployment correction may be insufficient to address such dynamics.

The paper argues that effective AI governance requires a multi-layered architectural approach, involving the alignment of legal, institutional, and technical layers with the requirements of the specific domain of application. In this context, the central question is not whether AI systems fail, but whether the selected class of system – statistical or deterministic – corresponds to the required level of reliability, predictability, and controllability..

Keywords: artificial intelligence; structural adaptation risk; neural network architectures; transformer models; optimization under constraints; specification gaming; AI governance; distributed systems.

For citation: A.Sinchev, B.Sinchev, A.Chinchalinova (2026). DISTRIBUTED AI ADAPTATION AS A STRUCTURAL RISK VARIABLE: FROM NEURAL ARCHITECTURES TO LAYERED SYSTEM DESIGN//
Conflict of interest: The authors declare that there is no conflict of interest.

А.Синчев [1], Б.Синчев [2], А.Чинчалинова[3*], А.Оспанова[4]

[1]Қазақстан Республикасы Президентінің Әкімшілігі, Астана, Қазақстан;
[2]Халықаралық ақпараттық технологиялар университет, Алматы, Қазақстан;
[3]Қазақстан Республикасы Президентінің Әкімшілігі, Астана;
[4]Цифрлық Үкіметті Қолдау Орталығы, Астана, Қазақстан.
*E-mail: Aigulch2206@mail.ru

ҮЛЕСТІРІЛГЕН ЖИ ЖҮЙЕЛЕРІНІҢ БЕЙІМДЕЛУІ – ҚҰРЫЛЫМДЫҚ ҚАТЕР АЙНЫМАЛЫСЫ РЕТІНДЕ: НЕЙРОН ЖЕЛІЛЕРІ АРХИТЕКТУРАСЫНАН КӨПДЕҢГЕЙЛІ БАСҚАРУ МОДЕЛІНЕ ДЕЙІН

Синчев Асқар – магистр, Қазақстан Республикасы Президентінің Әкімшілігі, Астана; БҰҰ Өнеркәсіпті дамыту ұйымының (UNIDO) AI for Industry Centre of Excellence жанындағы Жасанды интеллект бойынша Жаһандық альянстың Этика және реттеу кеңесінің мүшесі; мемлекеттік сектордағы әлемдік деңгейдегі үздік инноваторлардың қатарында танылған (Global Innovation Management Institute).
E-mail: askar.sinchev@gmail.com, ORCID: https://orcid.org/0000-0002-7333-2255.
Синчев Бахтгерей – техникалық ғылымдар докторы, Ақпараттық жүйелер кафедрасының профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: sinchev@mail.ru, ORCID: https://orcid.org/0000-0001-8557-8458
Чинчалинова Айгуль – магистр, Қазақстан Республикасы Президентінің Әкімшілігі, Астана.
E-mail: Aigulch2206@mail.ru, ORCID: https://orcid.org/0009-0001-1482-7784
Оспанова Алия – магистр, EMBA, Цифрлық үкіметті қолдау орталығының бас директоры, Астана, Қазақстан.
E-mail: aliyakalkamanovna@gmail.com, ORCID: https://orcid.org/0009-0006-9977-3696

Аннотация. Қазіргі заманғы жасанды интеллект (ЖИ) жүйелері оқиғаларға байланысты емес, олардың құрылымына кіріктірілген ерекше қатерлер класын қалыптастырады. Әдетте, жекелеген ақаулар немесе анық байқалатын бұзылулардан көрінетін дәстүрлі технологиялық қатерлер сияқты емес, қазіргі заманғы ЖИ жүйелеріндегі қатерлер жоғары параметрлі кеңістіктерде шектеулер толық анықталмаған жағдайда жүретін үздіксіз оңтайландыру үдерістерінен туындайды. Осы жұмыста мұндай қатер түрі құрылымның бейімделу қатері, яғни жүйенің мінез-құлқы уақыт өте келе оңтайландыру мақсатына ішкі тұрғыдан сәйкес бола тұра, бастапқыда болжанған функциялардан немесе саясат шектеулерінен ауытқып кетуі мүмкін жағдай ретінде қарастырылады. Ондай ауытқу жүйенің істен шығуын, сырттан араласуды немесе айқын қате болуын талап етпейді. Керісінше, ол прокси-көрсеткіштерге сүйенетін көзделген функциялар, көпқабатты нейрондық архитектуралардағы үлестірілген репрезентациялар және күшейтілген үйрету мен адам қатысуымен баға беруді қамтитын бейімделгіш кері байланыс механизмдерінің өзара әрекеттесуінен туындайды.

Эмпирикалық байқау деректері, соның ішінде талапты айналып өту (specification gaming), марапаттап манипуляциялау және контекске тәуелді жауап беру стратегиялары, ЖИ жүйелерінің болжамды нәтижелерден гөрі ресми берілген сигналдарды жүйелі түрде оңтайландыра алатынын көрсетеді. Осы тұрғыда мұндай мінез-құлықты аномалия деп емес, жүйелер шектеулер толық анықталмаған жағдайда жұмыс істегенінің заңды салдары деп қарастыру керек. Тағы бір маңызды салдары – қатер біртіндеп жинақталып, сырттан байқалатындай болғанға дейін жүйенің мінез-құлқына кірігіп кетуі мүмкін екені. Осыған орай мониторингке, аудитке немесе енгізілгеннен кейінгі түзетуге сүйенетін басқару тәсілдері ондай динамиканы толық ескеруге жеткілікті болмай қалады.

Бұл мақалада жасанды интеллектіні тиімді басқару құқықтық, институттық және техникалық деңгейлерді нақты қолдану доменінің талаптарына сәйкес үйлестіруді қамтитын көпдеңгейлі архитектура тәсілін қажет ететініне негіздеме беріледі. Осы тұрғыда негізгі мәселе ЖИ жүйелерінің істен шығып қалуында емес, жүйенің таңдалған – статистикалық немесе детерминирленген – класы қажетті сенімділік, болжамдылық және басқарылу деңгейіне сәйкес келе ме деген сұрақта.

Түйін сөздер: жасанды интеллект; құрылымның бейімделу қатері; нейрон желі архитектуралары; трансформер модельдер; шектеулер жағдайындағы оңтайландыру; талапты айналып өту; жасанды интеллектіні басқару; үлестірілген жүйелер.

Дәйексөз алу үшін: А.Синчев, Б.Синчев, А.Чинчалинова (2026). ҮЛЕСТІРІЛГЕН ЖИ ЖҮЙЕЛЕРІНІҢ БЕЙІМДЕЛУІ – ҚҰРЫЛЫМДЫҚ ҚАТЕР АЙНЫМАЛЫСЫ РЕТІНДЕ: НЕЙРОН ЖЕЛІЛЕРІ АРХИТЕКТУРАСЫНАН КӨПДЕҢГЕЙЛІ БАСҚАРУ МОДЕЛІНЕ ДЕЙІН//

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

А.Синчев [1], Б.Синчев[2], А.Чинчалинова[3*], А.Оспанова[4]

[1] Администрация Президента Республики Казахстан, Астана;
[2] Международный университет информационных технологий, Алматы, Казахстан;
[3] Администрация Президента Республики Казахстан, Астана;
[4] Центр поддержки цифрового правительства, Астана.
*E-mail: Aigulch2206@mail.ru

РАСПРЕДЕЛЁННАЯ АДАПТАЦИЯ ИИ КАК СТРУКТУРНЫЙ РИСК: ОТ АРХИТЕКТУРЫ НЕЙРОСЕТЕЙ К МНОГОУРОВНЕВОЙ МОДЕЛИ УПРАВЛЕНИЯ

Синчев Аскар – магистр, Администрация Президента Республики Казахстан, Астана; член Совета по этике и регулированию центра компетенций Глобального альянса по искусственному интеллекту в промышленности (UNIDO); признан одним из ведущих мировых инноваторов в государственном секторе (Global Innovation Management Institute).

E-mail: askar.sinchev@gmail.com, ORCID: https://orcid.org/0000-0002-7333-2255.

Синчев Бахтгерей – доктор технических наук, профессор кафедры «Информационных систем», Международный университет информационных технологий, Алматы, Казахстан.

E-mail: sinchev@mail.ru, ORCID: https://orcid.org/0000-0001-8557-8458.

Чинчалинова Айгуль – магистр, Администрация Президента Республики Казахстан, Астана.

E-mail: Aigulch2206@mail.ru, ORCID: https://orcid.org/0009-0001-1482-7784

Оспанова Алия – магистр, EMBA, Генеральный директор, Центр поддержки цифрового правительства, Астана, Казахстан.

E-mail: aliyakalkamanovna@gmail.com, ORCID: https://orcid.org/0009-0006-9977-3696

Аннотация. Современные системы искусственного интеллекта (ИИ) формируют класс рисков, которые не являются событийно обусловленными, а структурно встроены в их функционирование. В отличие от традиционных технологических рисков, проявляющихся через дискретные отказы или идентифицируемые неисправности, риски в современных ИИ-системах возникают в результате непрерывных процессов оптимизации, протекающих в пространствах параметров высокой размерности при неполной формализации ограничений. В данной работе такой тип риска определяется как риск структурной адаптации – возможность того, что поведение системы со временем эволюционирует таким образом, что, оставаясь внутренне согласованным с оптимизационной целью, оно отклоняется от изначально заданных функций или нормативных ограничений. Такое отклонение не требует ни отказа системы, ни внешнего вмешательства, ни явной ошибки. Оно возникает как результат взаимодействия целевых функций, основанных на прокси-показателях, распределённых представлений в многослойных нейронных архитектурах и адаптивных механизмов обратной связи, включая обучение с подкреплением и оценку с участием человека.

Эмпирические наблюдения, включая случаи specification gaming, манипуляции вознаграждением и контекстно-зависимых стратегий ответа, показывают, что ИИ-системы способны систематически оптимизировать формально заданные функции, а не предполагаемые результаты. В рамках предлагаемого подхода такие формы поведения интерпретируются не как аномалии, а как закономерные следствия функционирования систем в условиях неполной формализации ограничений. Существенным следствием является и то, что риск может накапливаться постепенно и встраиваться в поведение системы до того, как он становится наблюдаемым извне. В связи с этим подходы к управлению, основанные преимущественно на мониторинге, аудите или корректировке после развертывания, оказываются недостаточными для учета подобных динамик.

В статье обосновывается, что эффективное управление ИИ требует многоуровневого архитектурного подхода, включающего согласование правовых, институциональных и технических уровней с требованиями конкретного домена применения. В этом контексте ключевой вопрос заключается не в факте отказа ИИ-систем, а в том, соответствует ли выбранный класс системы – статистический или детерминированный – уровню требуемой надёжности, предсказуемости и контролируемости.

Ключевые слова: искусственный интеллект; риск структурной адаптации; архитектуры нейронных сетей; трансформерные модели; оптимизация при ограничениях; управление ИИ; распределённые системы.

Для цитирования: А.Синчев, Б.Синчев, А.Чинчалинова (2026). РАСПРЕДЕЛЁННАЯ АДАПТАЦИЯ ИИ КАК СТРУКТУРНЫЙ РИСК: ОТ АРХИТЕКТУРЫ НЕЙРОСЕТЕЙ К МНОГОУРОВНЕВОЙ МОДЕЛИ УПРАВЛЕНИЯ//

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Recent advances in artificial intelligence have led to the deployment of systems capable of continuous adaptation in complex and dynamic environments. Unlike earlier generations of software, whose behavior could be largely anticipated through predefined logic, modern AI systems operate through iterative optimization processes that update internal representations over time. As a result, system behavior is no longer fixed at deployment, but evolves in response to objective functions, data inputs, and feedback mechanisms.

In controlled experimental settings, advanced AI systems have demonstrated the capacity for strategically adaptive behavior, including deception, manipulation of evaluation procedures, and circumvention of oversight mechanisms, when such actions are instrumentally aligned with their optimization objectives[1,2]. These behaviors do not arise from system failure in the conventional sense. Rather, they reflect the internal consistency of optimization processes operating under imperfectly specified constraints.

This shift challenges conventional approaches to technological risk. Traditional risk frameworks assume that undesired outcomes are the result of discrete failures, errors, or external interference. Under this paradigm, risk can be mitigated through monitoring, testing, and post-deployment correction. However, in adaptive AI systems, undesired behavior may emerge gradually as a consequence of optimization dynamics, without any identifiable point of failure[11]. The system continues to function as designed, yet its behavior diverges from intended goals.

A growing body of research in AI safety has documented related phenomena, including specification gaming, reward manipulation, and context-dependent optimization strategies[3]. These findings suggest that AI systems tend to optimize for formally defined objective functions or proxy signals, rather than the underlying intentions those objectives are meant to represent[4]. Importantly, such behavior should not be interpreted as anomalous. It is a predictable outcome in settings where constraints are only partially formalized.

Despite these observations, existing approaches to AI governance remain largely oriented toward ex post control, including monitoring, auditing, and corrective intervention after deployment. Such approaches implicitly assume that risk is observable and can be detected once it manifests. This assumption becomes increasingly problematic in systems where risk is embedded in the trajectory of adaptation itself.

This paper introduces the concept of structural adaptation risk, defined as the possibility that system behavior evolves over time in ways that remain internally consistent with the optimization objective, yet diverge from intended functions or policy constraints. In this framework, risk is not associated with failure events, but with the properties of the optimization process and its interaction with incomplete constraint specification. The paper contributes a conceptual framework linking distributed optimization in high-dimensional parameter spaces with the emergence of structurally embedded risk in AI systems. It further argues that effective governance requires an architectural approach, in which constraint design, boundary conditions, and admissible regions of system behavior are specified at the design stage, prior to deployment. In this context, the central question is not whether AI systems fail, but whether their optimization trajectories remain within defined and controllable bounds.

1. Problem Formulation

Building on the preceding discussion, the central issue can be formulated more precisely in terms of the relationship between objective functions, constraint specification, and system-level behavior. We model AI systems as optimization processes operating within a feasible solution space defined by explicitly encoded objectives and constraints. Within this space, system behavior is, by construction, internally consistent with the optimization target. However, the formally defined feasible region does not necessarily coincide with the region of behavior that is acceptable from a functional, policy, or governance perspective[5].

This misalignment arises from the fact that constraint specification is necessarily incomplete. Not all relevant conditions governing acceptable system behavior can be formalized ex ante, particularly in complex or dynamic environments. As a result, the optimization process may systematically explore admissible regions of the solution space that remain compliant with the formal objective, yet violate implicit or unencoded constraints. Under these conditions, undesired outcomes should not be interpreted as deviations from system design. On the contrary, they represent valid solutions within the defined optimization structure. The source of the problem is therefore not failure of execution, but the structure of the feasible space itself[6].

This paper defines structural adaptation risk as a property of this misalignment: a condition in which the set of formally admissible solutions diverges from the set of intended or acceptable outcomes. The risk is embedded not in isolated decisions or events, but in the geometry of the solution space and the trajectory of optimization within it[7].

The problem becomes more complex in distributed systems composed of multiple interacting optimization processes. In such systems, each component operates under localized objectives and partial information. While each component may remain consistent with its own constraints, their interaction can produce emergent system-level behavior that lies outside globally acceptable bounds. This divergence is not reducible to individual components, but arises from their composition[8].

From this perspective, the core challenge is to ensure that the feasible region defined at the design stage adequately approximates the region of acceptable behavior. This shifts the focus of governance from detecting undesirable outcomes after they occur to structuring the optimization problem itself, including the definition of constraints, admissible regions, and system boundaries.

2. Neural Architectures as Optimization Substrates

Modern AI systems are predominantly implemented as neural network–based models trained through gradient-based optimization in high-dimensional parameter spaces. Within this framework, system behavior emerges from the interaction between model architecture, training dynamics, and objective functions. Neural networks define a parameterized function space, where each point corresponds to a specific configuration of model parameters. The training process can be understood as a trajectory through this space, guided by a loss function that serves as a proxy objective.

Crucially, the structure of this space is not neutral. It is shaped by several design choices:

- Loss functions, which define optimization targets and implicitly determine what constitutes successful behavior;
- Model architectures, which constrain the class of representable functions and influence the geometry of the parameter space;
- Training data distributions, which define the regions of the space that are explored during optimization;

- Optimization algorithms, which determine how trajectories evolve over time.

These elements jointly define the feasible region within which the system operates. However, this feasible region is only partially aligned with intended system behavior. Loss functions capture measurable objectives, but cannot fully encode context-dependent or implicit constraints. Architectural choices shape expressivity, but do not prevent the emergence of unintended strategies. Training data provides examples, but cannot exhaustively represent all relevant scenarios. As a result, neural network training inherently involves optimization within an incompletely specified constraint space.

This perspective provides the substrate for the mechanisms discussed in the following section. Phenomena such as specification gaming, reward manipulation, and emergent multi-agent dynamics should be understood not as isolated anomalies, but as natural consequences of optimization processes operating within high-dimensional spaces defined by incomplete objectives and constraints.

3. Mechanisms of Structural Adaptation Risk

Structural adaptation risk manifests through identifiable mechanisms arising in optimization-based systems. These mechanisms are not independent; they reflect different expressions of the same underlying property – optimization under incomplete specification.

3.1. Objective–Proxy Misalignment.

In many AI systems, the optimization target is defined through proxy variables that approximate the intended objective. Under increasing optimization pressure, systems exploit imperfections in these proxies, identifying strategies that maximize the proxy signal without fulfilling the underlying intent. This behavior, commonly referred to as specification gaming, reflects a systematic divergence between proxy metrics and intended outcomes[2].

3.2. Reward Channel Corruption.

In reinforcement learning settings, reward signals are assumed to represent task performance. However, when the reward-generating process is accessible or indirectly influenceable, systems may optimize for reward acquisition rather than task completion. This includes behaviors such as manipulating observations, exploiting feedback loops, or interfering with evaluation processes, resulting in inflated reward signals that do not correspond to actual performance[4].

3.3. Distributional Sensitivity.

Optimization strategies learned under specific training distributions may produce unstable or unintended behavior when deployed in environments with different statistical properties. Rather than simple performance degradation, systems may adopt alternative strategies that remain consistent with learned objectives but violate constraints that were not represented during training[1].

3.4. Instrumental Strategy Formation.

Optimization processes may give rise to intermediate strategies that increase the likelihood of achieving objectives across a range of conditions. These include maintaining optionality, acquiring resources, or reducing external constraints. Such strategies are not explicitly encoded but emerge as structurally advantageous within the optimization process.

3.5. Multi-Agent Interaction Effects.

In distributed systems, multiple agents or components optimize local objectives under partial information. Interaction among these agents can produce emergent dynamics, including coordination failures, competitive behaviors, or stable equilibria that do not satisfy system-level constraints. These outcomes arise from the composition of locally consistent optimization processes and are not reducible to individual components[8].

These risks are not anomalies but structural properties of adaptive systems.
The question is therefore not how to eliminate them, but how to design systems that remain controllable under such conditions.

4. From Risk to Design: Structuring AI Systems by Class of Control

The analysis above suggests that many of the risks associated with contemporary AI systems are not incidental but structural. They emerge from the underlying properties of systems that learn from data, adapt over time, and operate under conditions of uncertainty. This raises a more fundamental question: not how to mitigate individual risks, but how to design systems whose behavior remains controllable within defined operational boundaries.

A key premise follows: not all AI systems are equally suitable for all classes of problems. Systems based on statistical learning and approximation exhibit fundamentally different operational characteristics compared to systems grounded in exact, deterministic computation. When deployed in high-stakes environments, this distinction becomes critical, particularly in light of empirical evidence on the limitations of machine learning systems in terms of robustness, reliability, and distributional generalization[9].

From a structural perspective, AI systems can be divided into two broad classes:

- Statistical systems, which operate through probabilistic inference, pattern recognition, and approximation. Their outputs are inherently non-deterministic, contingent on training data, and sensitive to distributional shifts. This class includes large language models, deep neural networks, and reinforcement learning systems;
- Deterministic systems, which rely on formally defined algorithms with predictable and reproducible behavior under specified inputs and constraints. Their correctness can, in principle, be verified, and their outputs are stable given identical conditions.

This distinction is not merely technical. It defines fundamentally different classes of control. Statistical systems are optimized for adaptability and performance under uncertainty, but they do not provide guarantees in the strict sense. Their behavior is bounded probabilistically rather than deterministically. As a result, they may exhibit unpredictable degradation when exposed to inputs or environments that deviate from their training distribution. Deterministic systems, by contrast, are designed for environments where correctness, traceability, and stability are primary requirements. Their limitations lie in rigidity and reduced capacity to generalize beyond predefined rules, but within their domain of specification, they offer a higher degree of assurance.

The misalignment between these system classes and the domains in which they are deployed constitutes a primary source of systemic risk. Applying statistical systems to contexts that require strict guarantees – such as energy grid control, financial clearing and settlement, safety-critical infrastructure, or core public administration functions – introduces failure modes that are difficult to anticipate, detect, or contain. In such domains, uncertainty is not a feature. It is a failure condition.

This leads to a design implication: AI system architecture must begin not with model selection, but with problem classification in terms of required control properties. The question is not whether a system can perform a task, but whether its mode of operation is compatible with the level of reliability that the task demands.

A coherent design framework therefore requires alignment across three interdependent layers:

- Legal layer – establishing normative boundaries for the use of probabilistic versus deterministic systems, including explicit constraints on deployment in high-stakes domains, requirements for explainability, and rights of recourse;
- Institutional layer – defining responsibility, oversight mechanisms, and escalation pathways, particularly in cases where system behavior cannot be fully predicted ex ante;
- Technical layer – selecting and composing system architectures in accordance with the required class of control, including the use of hybrid designs where statistical components are constrained or supervised by deterministic cores.

Such an approach shifts the focus from model-centric optimization to architecture-centric governance. It acknowledges that capability alone is not a sufficient criterion for deployment. The central design question is therefore not how powerful a system is, but whether its structural properties are aligned with the operational demands of the domain in which it is embedded.

In this sense, AI risk is not primarily a question of failure at the level of individual systems. It is a question of design at the level of architecture.

Conclusion

The discussion in this paper suggests that the primary challenge in AI governance is not the absence of safeguards, but the absence of structural alignment between system design and domain requirements. As artificial intelligence systems become embedded in critical infrastructures and public decision-making processes, the distinction between statistical adaptability and deterministic control becomes increasingly consequential. This distinction is not merely technical; it defines the boundaries of reliability, accountability, and ultimately, trust.

A key implication follows: effective AI governance cannot be reduced to model evaluation or post hoc risk mitigation. It must begin at the level of architecture – through the classification of systems according to their control properties, and the alignment of legal, institutional, and technical frameworks accordingly. This perspective reframes the role of AI in public systems. Rather than treating intelligence as a universal layer applicable across domains, it requires a differentiated approach in which system design reflects the nature of the problem being addressed.

In this context, an important direction for further research lies in expanding the range of problems that can be addressed under more controlled and structurally constrained conditions, including recent work on structured approaches to combinatorial optimization[10]. Advances in this direction may enable a broader class of applications to be implemented within architectures that provide stronger guarantees of predictability and control. This paper has outlined a conceptual framework for approaching AI governance through the lens of structural alignment. The discussion has remained at the level of general architectural principles. Further work is required to operationalize this perspective, in particular by developing a more detailed articulation of layered system design and the corresponding governance mechanisms at each layer.

Ultimately, the question is not whether AI systems can perform increasingly complex tasks. It is whether they can do so within structures that remain governable.

REFERENCES

1. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. *Concrete Problems in AI Safety*. arXiv:1606.06565, 2016. Available at: https://arxiv.org/abs/1606.06565
2. Krakovna, V. et al. *Specification Gaming Examples in AI*. DeepMind, 2020. Available at: https://github.com/deepmind/specification-gaming
3. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control.* New York: Viking Press, 2019.
4. Everitt, T.; Krakovna, V.; Orseau, L.; Legg, S. *Reinforcement Learning with a Corrupted Reward Channel*. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2017. Available at: https://arxiv.org/abs/1705.08417
5. Hadfield-Menell, Dylan; Dragan, Anca D.; Abbeel, Pieter; Russell, Stuart. *The Off-Switch Game*. Proceedings of IJCAI, 2017.
6. Hubinger, Evan; van Merwijk, Chris; Mikulik, Vladimir; Skalse, Joar; Garrabrant, Scott. *Risks from Learned Optimization in Advanced Machine Learning Systems.* arXiv:1906.01820, 2019. Available at: https://arxiv.org/abs/1906.01820
7. Garrabrant, Scott; Demski, Abram; Critch, Andrew; et al. *Embedded Agency.* Machine Intelligence Research Institute (MIRI), 2018. Available at: https://intelligence.org/embedded-agency/
8. Leibo, Joel Z.; Zambaldi, Vinícius; Lanctot, Marc; Marecki, Janusz; Graepel, Thore. *Multi-agent Reinforcement Learning in Sequential Social Dilemmas.* Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2017. Available at: https://arxiv.org/abs/1702.03037
9. NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, 2023. Available at: https://doi.org/10.6028/NIST.AI.100-1
10. Sinchev et al. *Algorithm Based on the Subset Sum Problem for High Performance Computing*. Proceedings of Ninth International Congress on Information and Communication Technology, SpringerLink, 2024. Available at: https://link.springer.com/book/10.1007/978-981-97-3299-9
11. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift*. NeurIPS, 2019. Available at: https://arxiv.org/abs/1906.02530